2011

The Influence of Student and School Variables on Student Performance on the New Jersey Assessment of Skills and Knowledge in Grade 8

Maria A. Periera
Seton Hall University

THE INFLUENCE OF STUDENT AND SCHOOL VARIABLES ON STUDENT PERFORMANCE ON THE NEW JERSEY ASSESSMENT OF SKILLS AND KNOWLEDGE IN GRADE 8

BY

MARIA A. PEREIRA

Dissertation Committee

Christopher H. Tienken, Ed.D., Mentor
Daniel Gutmore, Ph.D.
Jeffrey Graber, Ed.D.
Amiot P. Michel, Ed.D.

Submitted in partial fulfillment
of the requirement for the Degree
Doctor of Education
Seton Hall University

2011

SETON HALL UNIVERSITY
COLLEGE OF EDUCATION AND HUMAN SERVICES
OFFICE OF GRADUATE STUDIES

APPROVAL FOR SUCCESSFUL DEFENSE

Doctoral Candidate, **Maria A. Pereira,** has successfully defended and made the required modifications to the text of the doctoral dissertation for the **Ed.D.** during this **Spring Semester 2010**.

DISSERTATION COMMITTEE

(please sign and date beside your name)

Mentor:
Dr. Christopher Tienken 1.20.2011

Committee Member:
Dr. Daniel Gutmore 1/20/2011

Committee Member:
Dr. Jeffrey Graber 1/20/11

Committee Member:
Dr. Amoit Michel 1-20-2011

External Reader:

The mentor and any other committee members who wish to review revisions will sign and date this document only when revisions have been completed. Please return this form to the Office of Graduate Studies, where it will be placed in the candidate's file and submit a copy with your final dissertation to be bound as page number two.

ABSTRACT

THE INFLUENCE OF STUDENT AND SCHOOL VARIABLES ON STUDENT PERFORMANCE ON THE NEW JERSEY ASSESSMENT OF SKILLS AND KNOWLEDGE IN GRADE 8

This study examined the strength and the direction of the relationships between student (i.e., socioeconomic status, attendance, and gender) and school variables (i.e., formative assessment usage and ASI classification) found in the extant literature to influence student achievement in language arts and mathematics. Analyses were conducted using simultaneous multiple regression models. All student data explored in this study pertained to 670 students in Grade 8 enrolled in four middle schools located in a suburban/urban central New Jersey community during thc 2008-2009 academic school year. The results of the study revealed each school produced a combination of site specific results and results common across sites regarding the strength of each independent variable to predict student achievement.

ACKNOWLEDGEMENTS

To an irreplaceable committee, it was a pleasure to work with people that provided such thorough and efficient feedback within what felt like mere minutes of receiving my work. Your expertise truly guided me through this process to help me produce a dissertation that I am extremely proud of.

To Dr. Tienken, I could write you an acknowledgement as long as this dissertation and it still would not be enough to capture how truly grateful I am for all the support, encouragement, and guidance that you have provided during this process. Although I would have been able to write a dissertation with another mentor, I am confident that it was your expertise and dedication that made it possible for me to produce the most polished, high-quality work that you knew I was capable of in such a timely fashion. You are the epitome of what a mentor, scholar, and teacher should be; I look forward to learning more from you in the future.

To Dr. Gutmore, from the moment you mentioned you were a structuralist in our first class I knew two things: I was going to learn a lot from you, and that we were going to get along just fine! Two years later and you did not let me down! Whether you were helping me with coursework, an independent study, or this document, you were always a true professional and will forever remain the model educator in my mind.

To Dr. Graber, thank you for agreeing to be part of this process with me on such short notice. Your encouraging words and friendly demeanor, along with your extensive knowledge base, was just the appropriate combination of what I needed to get me through the rough patches.

To Dr. Michel, you were a mentor to me long before you were even a member of my committee. Your work has been an invaluable resource to me. When I was lost many nights it was your dissertation that cleared the way for me. I reaped the benefits of your experiences and mastery of the topic of variables that influence student achievement and I was lucky to have the opportunity to work with you.

DEDICATION

This work is dedicated to my amazing husband, Tony, my wonderful parents, Anthony and Dora, and my supportive sisters Christina and Michelle. Thank you for always believing in me when I had lost faith in myself during this journey.

Tony, thank you for being my rock not only throughout this two year experience, but for the last ten years as well. You motivate me in ways that I can never express in words. The faith you had in me to finish this program was what kept me going throughout. Heck, I was ready to give up before I even had Chapter 1 completed! You traveled this journey with me, the highs and the lows, and I could not have asked for a better copilot. Your patience and understanding throughout this process has been unyielding and for that this degree is just as much yours as it is mine. I love you.

Baba and Mom, thank you for EVERYTHING! You raised me to truly believe wholeheartedly that I could have whatever title I desired before my name, whether that be Mrs., Dr., or even President. I am fortunate and eternally grateful to have parents that are the epitome of the American Dream. Thank you for providing me a life that made me want for nothing, but strive for everything. If I get to live a hundred lifetimes I could never repay you both for all that you have done. The very least I could do is dedicate this work to you. Just for you Baba, here it is in print—for the record…

Dr. Maria Antoinette LoGrande

I want to thank Christina and Michelle for reminding me to take a break once and a while and live in the moment. Now we can get back to our regular routines!

TABLE OF CONTENTS

List of Tables

List of Figures

Chapter I

INTRODUCTION

Background

No child left behind; four seemingly powerful, yet harmless words. This phrase is reminiscent of the armed forces of the United States and the promise to "leave no man behind." The No Child Left Behind (NCLB) Act education reform policy passed in 2001 and signed into law by President George W. Bush January 8, 2002 was "designed to improve student achievement and change the culture of America's schools" (NCLB A Desktop Reference, 2002, p.9). When President George W. Bush signed the NCLB Act into law, U.S. Secretary of Education Roderick Paige acknowledged that although many American schools did an adequate job of educating some of America's youth, NCLB marked the promise of "providing all our children with access to a high-quality education" (NCLB A Desktop Reference, 2002, p.9).

During his 2002 ceremonial signing of the NCLB Act at Hamilton High School in Ohio, President Bush announced "we are asking states to design accountability systems to show parents and teachers whether or not students can read and write and add and subtract" (Rogers, 2006, p.614). Instantaneously accountability became synonymous with test results. Instead of using the test "accountability system" as a diagnostic tool to assist educators in differentiating and driving academic instruction, tests became the primary indicator of a school's performance status (Rogers, 2006).

The NCLB Act mandates that all states focus on improving student academic standings while bridging the achievement gaps of all students. Four principles that steer the education reform policy include: (a) stronger accountability for results; (b) increased flexibility and local control; (c) expanded options for parents; and (d) an emphasis on teaching methods that have been proven to work (NCLB, 2001). To guarantee that states meet the required goal of one-hundred percent proficiency by the year 2014, the NCLB Act mandates that each state measure the adequate yearly progress (AYP) attained toward reaching this goal for all students in language arts and mathematics. Each state individually implements AYP targets or benchmarks, to ensure this goal is achieved by the year 2014. Districts that fail to meet AYP targets are held accountable under the NCLB Act stipulations.

In order to make decisions on the education reform necessary for the students of New Jersey to continue to display increased academic achievement levels, or lack thereof, the New Jersey Department of Education (NJDOE) uses the results from the NCLB Act required annual measurement of student achievement. The assessment currently used by New Jersey is the New Jersey Assessment of Skills and Knowledge (NJ ASK) administered in grades 3-8 for language arts and mathematics, and grades 4 and 8 for science content, to monitor the state's progress toward reaching AYP targets.

Education reform initiatives based on raising standardized test scores have inundated American classrooms with countless practice tests, various formative assessment tools, test preparation driven instruction, as well as constant curriculum revisions and modifications (Ryan, 2006). The recent implementation trend of formative and summative assessments in the classroom is a direct result of the quick-fix reaction by policymakers to adhere to NCLB Act requirements (Perie, Marion, & Gong, 2007; Plake, 2002; Sloane & Kelly, 2003).

The prevailing belief that the more students practice taking tests regardless of the tests' construct validity ultimately leads to increased levels of student achievement on state administrated tests is not fully supported empirically by methodologically sound studies, but far too common is the mindset of today's educational leaders and policymakers. Stiggins (2002) echoes this concern by identifying America as "a nation obsessed with the belief that the path to school improvement is paved with better, more frequent, and more intense standardized testing" (p.759). Although an insignificant amount of empirical research exists to support the measurable effectiveness or validity formative assessment instruments have on increased student achievement, these tools continue to flood 21st century classrooms (Dunn & Mulvenon, 2009b; Plake, 2002).

Statement of the Problem

A school's performance is identified and labeled by the state, policymakers, newspapers, and other media as either "successful" or "in need of improvement" primarily by state test results. In turn, school district leaders and administrators place great emphasis on state standardized test results to make what is believed to be "informed" decisions regarding future student placement and overall academic standings (Tienken, 2008a). This practice is commonplace "despite considerable evidence that high-stakes testing distorts teaching and does not give very stable information about school performance" (Dorn, 1998, p.2). Evidence indicates the NJ ASK and similar tests "have technical limitations and flaws that call into question the use of the results from those tests as high-stakes evaluative and decision-making tools" (Teinken, 2008a, p.48).

The state of New Jersey uses the District Factor Group (DFG) system for ranking the socioeconomic status of school districts. In a survey conducted for a descriptive study of the technical characteristics of the results from NJ ASK conducted by Tienken (2008a) to ascertain the predominate use of state standardized test results, evidence indicated that, regardless of a district's DFG, 98 percent of leaders that participated acknowledged using test results to make decisions. Additionally, roughly more than half of the leaders used the state test results as the only or the most important factor to make high-stakes decisions regarding student placement in remedial courses or to determine students' academic tracks.

The DE District in which I collected data does not analyze standardized language arts and mathematic assessment results jointly, but conversely, as if they are two separate entities. Because Grade 8 is considered a major academic transition period for students exiting middle school and entering high school, the DE District's administrative staff uses the Grade 8 NJ ASK results to track students for high school course placement. Although student achievement is monitored by the NJDOE and the NCLB mandates testing from Grades 3-8 and in Grade 11, due to the DE District's policies, I focused solely on Grade 8 student achievement in my study. For this study pseudonyms were used for the commercially produced formative assessment tool (FAT) and the publisher and vender of this education product (The Company).

New Jersey Core Curriculum Content Standards (NJCCCS) established benchmarks for determining what students should know and be able to do (NJDOE, 2006c). Statewide tests serve as a monitoring device to determine if curricula, funded programs, and classroom instruction are in alignment with the NJCCCS. The assumption is that if teachers are aligning lessons to the NJCCCS, then statewide test results should reflect this implementation. Prior to the implementation of the NJCCCS testing, Madaus (1988) prophetically stated "it is the testing, not the 'official' stated curriculum, that is increasingly determining what is taught, how it is taught, what is learned, and how it is learned" (p. 83).

Research reveals the problem: test results in actuality inform district leaders and administrators of the socioeconomic status (SES) factors and trends that are prevalent in the community rather than speak to the alignment of curricula to the NJCCCS. In New Jersey, there is a perfect Spearman Rho correlation between district level test scores and the SES of the community that the school serves (Tienken, 2008a).

No empirical literature exists that fully explains the relationship between student and school factors and NJ ASK scores in Grade 8. Furthermore, a review of the literature, both recent and past, pertaining to the effect and influence that formative assessment has upon student achievement lacks quantitative data to determine the efficiency of its use in classroom as a school reform tool (Dunn & Mulvenon, 2009a). Is the mere inclusion of formative assessment practices in the classroom a facilitator to student success despite other research-based variables known to influence achievement? It is essential to research and supplement what little empirical data is available to determine the correlation formative assessment practice and other school and student factors have to improving student achievement on the NJ ASK8.

Purpose of the Study

The purpose of this study was to determine the strength and the direction of the relationships between student and school variables found in the extant literature to influence student achievement and aggregate school NJ ASK scores in Grade 8 language arts and mathematics. By focusing on multiple school and student variables that significantly impact student achievement, this study aimed to produce research-based evidence to assist all stakeholders in public education regarding reform initiatives. Minimal conclusive empirical evidence exists regarding student and school variables and the influences of these variables on student achievement at the middle school level, specifically on Grade 8 state mandated tests of mathematics and language arts. Therefore this study could add to the limited body of existing literature to reshape public education.

Research Questions

Obtaining data from local school district student databases and the NJDOE, I attempted, through multiple regression, to determine the strength and direction of the relationships between student and school variables and academic performance on state mandated tests. This study was guided by the following overarching research question: What student and school variables, found in the extant literature explain the greatest variance in student achievement on the NJ ASK8 language arts and mathematics sections?

Research Question 1: What is the strength and direction of the relationship between the proficiency categorizations of students in Grade 8 measured by commercially produced standardized formative assessment in language arts and the proficiency categorizations of students in Grade 8 measured by the language arts portion of the NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Research Question 2: What is the strength and direction of the relationship between the proficiency categorizations of students in Grade 8 measured by commercially produced standardized formative assessment in mathematics and the proficiency categorizations of students in Grade 8 measured by the mathematics portion of the NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Research Question 3: What are the statistically significant student variables that explain the largest amount of variance in student language arts achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Research Question 4: What are the statistically significant student variables that explain the largest amount of variance in student mathematics achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Research Question 5: What are the statistically significant school variables that explain the largest amount of variance in student language arts achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Research Question 6: What are the statistically significant school variables that explain the largest amount of variance in student mathematics achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Null Hypotheses

Null Hypothesis 1: No statistically significant relationship exists between a commercially produced standardized formative assessment tool proficiency categorization and students' language arts proficiency categorization on the NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Null Hypothesis 2: No statistically significant relationship exists between a commercially produced standardized formative assessment tool proficiency categorization and students' mathematics proficiency categorization on the NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Null Hypothesis 3: There are no statistically significant, research demonstrated, student variables that predict student language arts achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Null Hypothesis 4: There are no statistically significant, research demonstrated, student variables that predict student mathematics achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Null Hypothesis 5: There are no statistically significant, research demonstrated, school variables that predict student language arts achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Null Hypothesis 6: There are no statistically significant, research demonstrated, school variables that predict student mathematics achievement as measured by the state

mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Significance of the Study

Accountability is in the forefront of the education reform movement. As a result, school districts across the United States have spent millions of dollars on formative assessment tools in hope of shaping students into productive globally-prepared citizens of the future. The costs per pupil, per school year ranges anywhere from $19 to upwards of $54 on formative, interim, and summative assessments (APQC, 2005; Piton Foundation & Donnell-Kay Foundation, 2007). This study could benefit school administrators, educators, curriculum leaders, parents, and school boards as well as education researchers in determining what impact, if any, formative assessment has on student achievement and how to best spend scarce resources. The study of these uncharted waters will also add to the current, limited empirical evidence available in the literature to either support or challenge the positive implications associated with commercially prepared formative assessment products vended to schools.

The value and information gained from state test results is only as significant as the observer desires it to be. The American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999) expressed in a statement that:

> Although not all tests are well developed nor are all testing practices wise and beneficial, there is extensive evidence documenting the effectiveness of well-constructed tests for uses supported by validity evidence. The proper use of tests can result in wiser decisions about individuals and programs than would be the

> case without their use and also can provide a route to broader and more equitable access to education and employment. The improper use of tests, however, can cause considerable harm to test-takers and other parties affected by test-based decisions. (p.1)

It is crucial that district leaders and administrators acknowledge the irreversible potential harm of exclusively using state test results for high-stakes purpose and the misfortune this may inflict upon particular students of the community.

Limitations

"Non-experimental research is frequently an important and appropriate mode of research in education" (Johnson, 2001, p.3) due largely in part to the inability to perform randomized experiments and quasi-experiments. I conducted a non-experimental, cross-sectional, explanatory study. This correlational study only collected data from one point in time. Under the auspices of Johnson (2001) an explanatory study must meet the following criteria: (a) Were the researchers trying to develop or test a theory about a phenomenon to explain "how" and "why" it operates? (b) Were the researchers trying to explain how the phenomenon operates by identifying the causal factors that produce change in it? (p.9).

The data for this study was collected from one district labeled with a DE DFG. Generalizations cannot be made that similar results would prevail to some extent within the same DFG or in areas associated with a wealthier DFG. It is important to note that although the schools in the district of the study are grouped in a DE DFG, not all schools represent the DE category. For example, some schools are more representative of a CD or GH DFG in terms of SES characteristics. This study focused on one commercially produced standardized formative assessment product and only variables identified previously in the literature that influence student achievement (e.g. socioeconomic status, school characteristics, etc.).

Effective learning styles vary among students, thus affecting how learners approach test taking situations differently (Boyle, Duffy & Dunleavy, 2003). These inherent differences, along with student self motivation levels, personal learning styles, and belief systems can threaten the validity of the study. A refusal to partake in the testing procedures and formative pre and post assessments may have affected the results of the individual student's performance. This study only focused on Grade 8 and not on high school or elementary grade levels. In this DE district the total number of irregularities reported on the New Jersey Assessment of Skills and Knowledge (NJ ASK) for grades 6, 7, and 8 during the time of the study was ten. This study does not compare the scores of regular education students with special education students, but rather the formative assessment results with those of the NJ ASK summative scores for the same student to track individual increased student achievement levels.

Garson (2010) explained that "multiple analysis of variance (MANOVA) is used to see the main and interaction effects of categorical variables on multiple dependent interval variables" (p.1). Also the MANOVA approach "uses one or more categorical independents as predictors" (p.1). Because "MANOVA tests the differences in the centroid (vector) of means of the multiple interval dependents for various categories of the multiple interval dependents for various categories of the independent (s)" (p.1), it would have been useful to use this statistical method for my study, however the district leadership considers the Grade 8 language arts and mathematics NJ ASK results to be mutual exclusive, instead of correlated, which they are with a Pearson correlation coefficient of 0.74 (NJDOE, 2009). The district leadership uses the results from the language portion to make student placement decisions in language arts, and the math results to make student decisions in math. Therefore, I conducted a series of regression analyses using single output variables to investigate relationships between multiple independent variables and individual dependent variables.

Delimitations

Data was retrieved for Grade 8 students in language arts and mathematics across four middle schools located in a New Jersey DE District. The primary data sources were the FAT pretest and posttest assessments as well as NJ ASK8 scores. The examined test scores were from students enrolled in the DE District during the 2008-2009 school year.

Data was analyzed by building and not aggregated to the district level. The findings of this study will assist other school districts that are contemplating incorporating a formative assessment tool such as FAT for use in their school district by providing documented empirical results. The DE District is one of the largest in the state of New Jersey and is comprised of 24 schools with a well diverse population (Local Government Budget Review, 2001). Analysis developed via this study would benefit an array of both large and small New Jersey school districts as well as out of state school districts.

It is imperative to emphasize the impossibility of including all potential variables that may influence student achievement into this data collection (e.g., mandated instructional programs, instruction delivery strategies, professional development implications, and technology infusion).

Definition of Terms

Accountability: In accordance with NCLB each state must devise and implement a plan that details how and under what timeframe adequate yearly progress targets will be set and eventually met to increase student achievement levels.

Achievement Gap: As defined in popular literature, is the difference in student achievement between various groups of students (e.g., White and Black; rich and poor).

Adequate Yearly Progress (AYP): NCLB mandates that each state measure the progress made toward reaching the goal of one-hundred percent proficiency for all students in language arts and mathematics. Each state implements targets or benchmarks, to ensure this goal is achieved by the year 2014. Districts that fail to meet AYP targets are held accountable.

District Factor Group (DFG): The state of New Jersey uses the District Factor Group system for ranking the socioeconomic status of school districts. (see Chapters II and III for more information regarding DFG)

FAT (pseudonym for commercially produced Formative Assessment Tool): FAT is a web-based technology formative assessment tool aligned with individual state's core curriculum content standards and standardized state assessments. This tool is designed to help diagnose strengths and weaknesses of individual students on benchmarked assessments which provide teachers with immediate data which should subsequently be used to drive classroom instruction.

Formative Assessment: For the purpose of this study the definition formulated by Perie, Marion, and Gong (2007) was used. "Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes" (p.1).

Interim Assessment: For the purpose of this study the definition formulated by Perie, Marion, and Gong (2007) was used. Interim assessments are "the assessments that fall between formative and summative assessment, including medium-scale, medium-cycle assessments currently in wide use" (p.1).

Middle School: For the purpose of this study, middle school refers to the educational grade levels of 6, 7, and 8.

New Jersey Assessment of Knowledge and Skills (NJ ASK): NCLB requires the annual measurement of student achievement. The assessment currently used by New Jersey is the NJ ASK which is administered in grades 3-8 for language arts, mathematics, and science content areas to monitor the state's progress toward reaching AYP targets.

New Jersey Core Curriculum Content Standards (NJCCCS): The NJCCCS adopted in 1996, identify what students are expected to know and be capable of doing in nine different content areas at the conclusion of a 13 year public education.

No Child Left Behind (NCLB): Congress passed the NCLB education reform policy in 2001 which President George W. Bush later signed into law January 8, 2002. NCLB mandates that all states focus on improving student academic standings while bridging the achievement gaps of all students. States are required to meet the goal of one-hundred percent proficiency by the year 2014.

Proficiency Levels: The NJ ASK determines the performance level descriptors for all 3-8 language arts, mathematic, and science assessments. For all content areas, a scaled score between 100-199 falls in the partially proficient range, 200-249 falls in the proficient range and 250-300 falls in the advanced proficient range.

Student Achievement: For the purpose of this study, student achievement occurs at the point in which a student's scaled score falls in the "proficient" range on the NJ ASK assessment.

Summative Assessment: For the purpose of this study the definition formulated by Perie, Marion, and Gong (2007) was used. "Summative assessments are generally given one time at the end of some unit of time such as the semester or school year to evaluate students' performance against a defined set of content standards" (p.1).

Chapter II

REVIEW OF LITERATURE

Introduction

The purpose of this study was to determine the strength and the direction of the relationships between student, school, and teacher variables found in the extant literature to influence student achievement and aggregate district student NJ ASK scores in Grade 8 language arts and mathematics. The review of literature was comprised of the proceeding sections: Statewide Standardized Testing, High-Stakes Testing, Socioeconomic Status and Student Achievement, Attendance and Student Achievement, Gender and Student Achievement, Formative Assessment and Student Achievement, Interim Assessments, and Teacher Variables and Student Achievement.

The purpose of the review was to identify empirical studies that attempt to determine what statistical significance, if any, student, school, and teacher variables have on student achievement in Grades 8 as measured by the NJ ASK tests of language arts and mathematics. The intent is to inform education leaders, researchers, and policymakers about the present evidence regarding student achievement predictors.

Literature Search Procedures

The literature reviewed for this chapter was accessed via online databases including EBSCOhost, ProQuest, ERIC, JSTOR, and Academic Search Premier as well as online and print editions of peer-reviewed educational journals. Each section of reviewed literature includes experimental, quasi-experimental, meta-analysis, and/or non-experimental treatment/control groups studies. In order to effectively and systemically "present results of similar studies, to relate the present study to the ongoing dialogue in the literature, and to provide framework for comparing the results with other studies" (Creswell, 1994, p.37), I followed the framework for scholarly literature reviews developed by Boote and Beile (2005).

Methodological Issues in Studies of Predictors on Student Achievement

When reviewing the literature, I encountered several issues regarding the three main variables, student, school, and teachers, associated with predicting student achievement on state standardized tests. The research related to each of the variables suffered from various methodological issues: (a) the lack of experimental studies, therefore placing a heavy reliance on correlational designs; (b) the consistent absence of the reporting of experimental effect sizes; (c) the reporting of varying, mixed results that were gathered using the same data; and (d) the lack of clarify on terms used.

In an attempt to confront the aforementioned issues, I chose to include as many pertinent experimental studies as possible, but also non-experimental and quasi-experimental research to fuel my literature review. Johnson (2001) clarified best when he wrote:

> Although the strongest designs for studying cause and effect are the various randomized experiments, the fact remains that educational researchers are often faced with the situation in which neither a randomized experiment nor a quasi-experiment (with a manipulated independent variable) is feasible. (p.3)

Johnson affirmed that "nonexperimental research is frequently an important and appropriate mode of research in education" (p. 3), and therefore it was effectively incorporated in my literature review.

To overcome the frequent lack of efficient effect size reporting within literature reviewed, the I reported study effect sizes were calculated by hand, when the data and required information was made available by the researcher (s). By calculating effect size and using Cohen's (1977) level of significance (0.00-0.25 = small; up to 0.50 = moderate; 1.00+ is large), at times I was able to identify weaknesses and flaws in the researcher(s) results as to the accuracy of the level of significance purported.

In many studies the same terms were used with different definitions. Whenever the possibility existed that there was confusion regarding the usage of a term, I provided a synthesized definition from the literature. For example, there is no clear, concise, widely accepted definition of the term "formative assessment" (Black & Wiliam, 1998; Dunn & Mulvenon, 2009b; Leung & Mohan, 2004). Due to the issues with terminology, searching for the literature relating to "formative assessment" proved a challenging feat. Many of the positive aspects associated with formative assessment classroom implementation by educators and administrators today are grounded in the work *Assessment and Classroom Learning* published by Black and Wiliam in 1998. Dunn and Mulvenon (2009a) emphasized the profuse references of the Black and Wiliam piece by noting that "the Social Science Index indicates that it has been referred to in scholarly journals 194 times" (p. 5). Hence, I used the Black and Wiliam (1998) seminal work as a starting point when reviewing literature pertaining to that topic.

Inclusion and Exclusion Criteria for Literature Review

Studies that met the following criteria were included in this review:

1. Used experimental, quasi-experimental, non-experimental with control groups, or another design that would be considered at least causal-comparative.

2. Peer-reviewed, dissertations, or government report.
3. Report at least statistical significance.
4. Published within the last 30 years unless considered seminal work and thus older.
5. Included the use of formative or interim assessment as one intervention.
6. Any literature, that meets the above design criteria, found in a report from a governmental body advocating the use of formative or interim assessment.

The review emanates from one of the most heavily cited works on formative assessment (Black & Wiliam, 1998). As a baseline for their evaluation of formative assessment used in classrooms, Black and Wiliam (1998) reviewed one article by Natriello (1987) and one article by Crooks (1988). Between the works of Natriello and Crooks, Black and Wiliam collected 681 references effectively demonstrating their criteria of "formative assessment", even if it went by a different term, such as "classroom evaluation". This was later narrowed down to about 250 publications, that according to Black and Wiliam, were "sufficiently important to require reading in full" (p. 2). Although Black and Wiliam found relevance in 250 publications related to their purpose, included in this dissertation are only the eight studies labeled as "Examples in Evidence/ Classroom Experience" (p. 3). Black and Wiliam expressed the purpose of including these particular studies as "aims to secure evidence about the effects of formative assessment" (p.3). The eight studies showcased by Black and Wiliam to emphasize the positive implications of formative assessment on student achievement include a vast array of activities that involve components of feedback and modifications in learning/teaching practices conducted by student and/or teacher.

The types of formative assessment used in the eight studies varied and included feedback and techniques placing emphasis on goal orientation, self-perception, self-assessment, self-evaluation, and frequent testing. Literature from both the 20th and 21st centuries was included to showcase current trends in formative assessment in the field of education. I am interested in reviewing and comparing the various "traditional" formative assessment practices (i.e. paper and pen, teacher feedback) using studies prior to the 21st century (specifically 1986-1997) and study effect size with recent web-based/technology infused formative assessment programs (specifically 1999-2007) and the implication these modern tools have on student achievement. However, excluded from this review are additional studies, other than the eight referenced by Black and Wiliam, discussing other types and methods of formative assessments available. The primary focus of the study is on web-based/technology infused formative assessment programs, however due to the influential work of Black and Wiliam, I included the eight signature studies not involving web-based/technology infused formative assessment tools for the above mentioned purposes.

This literature review is also comprised of five studies identified by Dunn and Mulvenon (2009a) as "educational technology literature" (p.8). I chose to focus on these studies because they were highlighted by Dunn and Mulvenon as suffering from methodological issues. All web-based/technology included literature studies were conducted between 1999 and 2007, and therefore much more recent and indicative of a 21st century classroom learning experience. Four of the five web-based/technology studies included by Dunn and Mulvenon were conducted at the college level. Although my research is on the middle school level, for consistency purposes I chose to focus on the five web-based/technology programs selected by Dunn and Mulvenon, for similar inclusion reasons mentioned previously for the "traditional" formative assessment studies. There are similarities between FAT, the formative assessment technology tool used in the study, and the programs used in the five referenced studies (i.e. web-based, pretest/posttest, customized reports, individual assessments).

Experimental, quasi-experimental, and non-experimental research included pertained exclusively to the specific student (i.e., socioeconomic, attendance, gender), school (formative assessment), and teacher (educational attainment) variables emphasized in my study. I included the historical seminal works on the topics and then empirical research from 1997 to 2010. Information regarding other studies on variables that may affect student achievement other than the aforementioned were excluded from the literature review solely because of their unrelated relevance to my study. Although the goal effect size when developing an intervention is 0.30 or larger in educational studies (Cohen, 1977), this literature review includes studies in which the effect sizes are insignificant or not reported at all, for the sole purpose of highlighting the weaknesses of that particular study.

Review of Literature Topics

Statewide Standardized Testing

In *The Principles of Scientific Management*, industrial analyst Frederick Taylor (1911) recounted his experiences at the Bethlehem Steel Company and the work he did with the scientific management element of "task idea". In an effort to demonstrate how scientific management was much more efficient than the current plan, Taylor studied a group of 75 pig iron laborers as they worked.

Taylor (1911) recalled "it was our duty to see that the 80,000 tons of pig iron was loaded on to the cars at the rate of 47 tons per man per day, in place of 12 ½ tons, at which rate the work was being done" (pp. 42-43). Taylor continued by explaining that the intention was not to cause an uprising of disgruntled employees or to have the men go on strike, but to give them the incentive to increase productivity. Through a series of experiments, Taylor determined the exact amount of rest a man that needed to move 47 tons of pig iron a day would require in order to avoid over exhaustion. Once that was established Taylor proceeded to find "the proper workman to begin with" (p.43). In the selection process, Taylor first had to identify men that would be physically capable of hauling 47 tons of pig iron in the first place. At that point four of the 75 men were selected. Taylor then conducted background investigations to learn about their character, ambitions, and habits of each of the four men in order to identify the one most suited for the incentive program.

Taylor (1911) explained to Schmidt, the man chosen for the program, that he would get $1.85 a day, a significant raise from his current rate of $1.15 a day, if he was willing to prove that he was "a high-priced man" (p. 45). By informing Schmidt of exactly what to do, when to do it, and for how long to do it, he was able to increase productivity from moving merely 12 ½ tons of pig iron to an astounding 47 ½ tons in one day, as well as increasing his daily earnings.

Taylor's (1911) scientific management approach was different than previous practices that attempted to motivate workers with the "initiative and incentive" method. What the previous methods lacked that Taylor incorporated in his model was informing the worker exactly what he must do in order to reach the goal rather than have the worker be responsible of figuring out how to do it himself.

Frederick Taylor's (1911) "scientific management" approach allowed for organizations to run on maximum efficiency by accentuating people's performance while working to get the most out of every second spent on the job. Rather than focusing on training someone to fit into a particular mode or enticing them with the allure of change, Taylor believed that people should be placed in the right roles and relationships solely based on the efficiency they bring to that position. At the commencement of World War I, the United States military subscribed to Taylor's industrial workplace efficiency tactics to classify recruits and volunteers.

In an effort to systemically assign volunteers and recruits to successful positions within the army ranks during World War I, officials consulted the American Psychological Association. The result was the advent of the standardized Army Alpha Tests, which included a subset of ten different tests, contrived to "discriminate among test-takers with respect to their intellectual abilities" (Popham, 2001, p.42). Men preparing for combat were administered the Army Alpha Test and the results generated determined the placement of these men. Officer training was reserved for those that ranked high on the Army Alpha Test, whereas lower rankers were assigned positions on the battlefields (Solley, 2007).

In the decades following World War I, "the number and variety of standardized test had increased exponentially and there was almost no sector of the U.S. society untouched by the standardized testing movement" (Kennedy, 2003, p.2). Achievement tests quickly emerged that replicated the Army Alpha Test. A prime example is the ever-present Scholastic Aptitude Test (SAT), currently taken by more than two million college-bound students yearly (College Board, 2010), due to the fact that many United States universities require this as part of their admissions criteria.

The landmark Eight-Year Study conducted by the Progressive Education Association (PEA) between 1930 to 1942 was an experimental project that advocated school curricula redesign in an effort to shift focus away from the college entrance dominated graduation requirements. The PEA recognized that only one out of every six American students in the late 1920s actually attended college upon high school graduation, however high school coursework was predominately comprised of college preparation programs.

In order to meet the needs of both college and non-college bound students, the PEA launched an experimental project that would help determine if a unified core curriculum would prove fruitful for all involved parties regardless of future educational goals. Thirty high schools and 250 colleges participated in the Eight-Year Study. School curricula were not the only area that was affected by the Progressive school approach, staff development, assessment practices, and student guidance procedures were also included in the redesign experiment.

The Eight-Year Study found: (a) the graduates of the 30 schools were not handicapped in their college work; (b) departures from the prescribed pattern of subject and units did not lessen the student's readiness for the responsibilities of college; and (c) students from the participating schools which made most fundamental curriculum revision achieved in college distinctly higher standing than that of students of equal ability with whom they were compared. Although the study produced favorable outcomes for the effectiveness of the Progressive style approach to education it is believed that with the advent of World War II, the Eight-Year Study and the positive implications associated with it unfortunately fell by the wayside.

In an effort to fight the "war on poverty", which subsequently quelled civil right tension, President Johnson's Administration vowed "the ending of discrimination against nonwhites, citing data that underscored the differences in the status of white and nonwhite Americans in education, employment, health care, and housing" (Amaker, 1988, pp. 19-20). As a result of Johnson's commitment the following acts were secured and passed into congress, the Civil Rights Act (1964), the Voting Act (1965), the Elementary and Secondary Education Act (1965) (ESEA), the Age Discrimination in Employment Act (1967), and the Fair Housing Act (1968) (Amaker, 1988).

The ESEA (1965) focused on supporting "local educational agencies serving area with concentrations of children from low-income families to expand and improve their educational programs by various means (including preschool programs) which contribute to meeting the special educational needs of educationally deprived children" (ESEA, 1965). During the first year of its enactment, ESEA awarded 960 million dollars to schools that served large numbers of disadvantaged, poverty-stricken, and minority families (Brookes & Pakes, 1993). Concerned with the prevailing evidence that vast discrepancies existed among the achievement levels of affluent and disenfranchised students, the Title 1 Program served to provide resources and programs that would assist in creating equal learning opportunities for all students of different socioeconomic backgrounds. The Title 1 Program provides federal funds to schools and local educational agencies with large enrollments of poor students in order to assist with meeting the academic needs of these students. As continuous reauthorizations of the act evolved, the amount of funding steadily increased, and by the early 1990s funds exceeded six billon dollars a year (Stringfield, 1991).

In order to determine the effectiveness of the funded programs and resources allocated via ESEA and Title 1 provisions on student achievement, an evaluation procedure was required (Solley, 2007). Policymakers and education leaders looked to the past efficiency model for testing large numbers of subjects and hence modified and adopted the Army Alpha Test for student academic use. Although the criterion-referenced tests available during the mid-1960s and 1970s "bore no direct relationship to the skills and knowledge being promoted by any particular ESEA program" (Solley, 2007, p.32) the use of these ineffectual measurement techniques continued. Using standardized achievement tests to assess the value of programs and resources lead to the assumption that the same measurement tool could evaluate the learning of children (Solley, 2007). Today, the inundation of standardized achievement tests in American classrooms and mandatory accountability evaluation reports can be linked directly and indirectly to the ESEA of 1965 (Brooks & Pakes, 1993; Kennedy, 2003; Solley, 2007).

Under President Reagan's Administration, Terrel Bell as Secretary of Education was required to oversee the disbanding of the Department of Education. Due to legal issues this never occurred, however Reagan was still set to gain politically. In an effort to link the economic hardships of the 1980s with the current state of the educational system, Bell formed The National Commission on Excellence in Education. The *A Nation At Risk* (USDOE, 1983) report was riddled with rhetoric that claimed American schools were in dire shape and needed immediate attention considering our economy and national security rested on the reform of the educational system. The goal of the Reagan Administration was to win the education debate against the Democratic Party, and at no expense, the *Nation at Risk* report solidified this. As a result of trends in educational reform, in addition to the influx of ideology and rhetoric published in performance reports, the increase in administration of mandatory testing was condoned by the public and viewed as an easy response to the growing concerns regarding student achievement (Archbald & Porter, 1990).

By the late 1980s and early 1990s there was an acceleration in the number of tests that were administered to students, due in part to the fact that many states now mandated testing students at multiple stages throughout the K-12 grade levels (Archbald & Porter, 1990). In 1990 Archbald and Porter reported:

> The National Center for Fair and Open Testing estimated K-12 students in the U.S. take around 100 million standardized tests, about an average of 2.5 standardized tests per student per year (this includes state and district testing). The National Commission on Testing and Public Policy estimates standardized testing costs between $700 million to $900 million yearly in purchasing costs and administration time, or about $17 to $22 per student per year. (p.12)

Prior to the NCLB Act enacted by the Bush administration in 2002, President Clinton's Goals 2000: Educate America Act (P.L. 103-227) of 1994 was the dominating federal education reform initiative. To ensure all students obtained high levels of academic achievement the Goals 2000: Educate America Act provided states with resources and a framework for education reform. A focal point of the act was the call for voluntary testing in grades four, eight, and twelve in order to provide evidence that students were making academic progress and meeting world-class academic standards.

By 2002, the NCLB Act had made statewide standardized testing of language arts and mathematics in grades 3-8 and one year in high school mandatory. Incremental and consistent growth in standardized test results is something that policymakers must see concrete evidence of in all school districts across the United States. As a result of the NCLB Act accountability expectations, all 50 states have implemented a form of standardized testing practices which subject students to yearly standardized assessments beginning in Grade 3. It is not uncommon for students to be tested as early as first and second grade with commercially prepared standardized tests in many school districts around the United States (Solley, 2007) as a precursor for the federally and state mandated exams beginning in Grade 3.

On February 17, 2009, President Obama signed into law the American Recovery and Reinvestment Act (ARRA). ARRA serves as a "response to a crisis unlike any since the Great Depression, and includes measures to modernize our nation's infrastructure, enhance energy independence, expand educational opportunities, preserve and improve affordable health care, provide tax relief, and protect those in greatest need" (NJDOE, 2010c). With effective teaching and learning at the core of the ARRA, reform elements include: (a) standards and assessments; (b) excellent teaching and leadership; (c) data systems; and (d) struggling schools (NJDOE, 2010c). Race to the Top (RTTT) was a 4.35 billion dollar competitive grant that was made available to school districts and local educational agencies (LEAs) via the ARRA during the 2009-2010 school year. RTTT was the largest education grant program in terms of dollars ever offered in the United

States. All school districts and LEAs were presented with the opportunity to receive RTTT funding by completing and submitting an application.

Unbeknownst to some, RTTT funding comes with strings attached. When submitting an application it was essential that district leaders were cognizant of the ramification this decision could have had on the future of their students. Can a price tag be placed upon a well-balanced human being? Districts and LEAs that applied to the grant can expect to: (a) increase standardized test preparation even more in both mathematics and language arts; (b) reduce the number of "electives" offered and increase the number of STEM (science, technology, engineering, and mathematics) courses available; and (c) create and implement "a common set of k-12 standards...that are supported by evidence that they are internationally benchmarked and build toward career readiness by the time of high school graduation" (NJDOE, 2010c), among many other things.

Zhao (2009) and others argue students need to grow and discover their individual talents, however with continuous pressure tied directly to "core" academic achievement (mathematics, language arts, and science), other programs that foster creativity are quickly being disbanded. At this rate there is a possibility that by the end of their high school education, students will not just enter college unprepared and one dimensional, but will only be accustomed to practicing and taking standardized assessments. Unfortunately, students will grow to lack the ingenuity to be 21st century explorers.

As standardized achievement tests are used more frequently today to assess students and used to make high-stake decisions, the similarities to the Army Alpha Test are not overlooked. Although the acronyms of standardized achievement test titles are ever-changing, one constant is the purpose, which remains to compare test-takers' scores "to a pre-determined norm group to discriminate among them and determine rank" (Solley, 2007, p.33). With no evidence of a lull in standardized achievement test administration on the horizon, the ominous message lamented by Kohn (2000) warning "standardized testing has swelled and mutated, like a creature in one of those old horror movies, to the point that it now threatens to swallow our students" (p.1) is quite fitting as 2014 draws closer.

High-Stakes Testing

Although a universally agreed upon definition is nonexistent, one certainty of the phrase "high-stakes testing" is that it is controversially regarded in the field of education today (Amrein & Berliner, 2002a,b; Amrein-Beardsley & Berliner, 2003; Braun, 2004; Carnoy & Loeb, 2002; Marchant, 2004; Marchant, Paulson, & Shunk, 2006; Nichols, Glass, & Berliner, 2006; Raymond & Hanushek, 2003; Rosenshine, 2003; Solley, 2007). According to the American Educational Research Association (as cited in Marchant, 2004, p.2) high-stakes testing requires:

> Many states and school districts mandate testing programs to gather data about student achievement over time and to hold schools and students accountable. Certain uses of achievement tests results are termed "high stakes" if they carry serious consequences for students or educators. Schools may be judged according to the school-wide average scores for their students. High school-wide scores may bring public praise or financial rewards; low scores may bring public embarrassment or heavy sanctions. For individual students, high scores may bring a special diploma attesting to exceptional academic accomplishment; low

> scores may result in students being held back in grade or denied a high school diploma.

Tienken (2008a) reported a definition from the existing literature on the subject:

> Three conditions must be present for a test or testing program to be considered high-stakes: (a) a significant consequence related to individual student's performance; (b) the test results must be the basis for the evaluation of quality and success of school districts; and (c) the test results are must be the basis for the evaluation of quality and success of individual teachers (Madaus, 1988, Popham, 2001). (p. 50)

Considering Tienken's observations, it is no coincidence that the standardized achievement tests administered by all states under the NCLB Act evaluation mandates are considered high-stakes tests (Dorn, 1998; Marchant, 2004; Popham, 1999; Rothstein, 2009; Solley, 2007; Stiggins, 2002). The consequences of high-stakes testing results affect more than AYP targets and local funding allocations; in fact the most valuable asset, American students, suffer the most from this backlash.

Marchant (2004) summarized that the guidelines for high-stake testing efforts issued forth by the 1999 Standards for Educational and Psychological Testing included:

> Protection against high-stakes decisions based on a single test, full disclosure of likely negative consequences of high-stakes testing programs, alignment of the test and the curriculum, opportunities for remediation for those who fail, appropriate attention to language differences and disabilities. (p.2)

Even with that knowledge, it is customary for administrators and education leaders to use, year after year, the results of standardized tests to determine whether students are retained, promoted, graduate, enrolled in remedial courses, even accepted into colleges (Marchant, 2004; Tienken, 2008a). This fact was further enforced in a survey conducted by Tienken (2008a) in which approximately 55 percent of questioned New Jersey education leaders acknowledged that high-stakes decisions to enroll students in courses that require basic skill instruction and acquisition were based predominately "on state test results" (p. 56). Furthermore, 98 percent of the same surveyed leaders admitted to using high-stake test results to make decisions, including curricula evaluations (Tienken, 2008a).

Studies have been conducted that examine how high-stakes testing has influenced student motivation (Amrein & Berliner, 2003; Good & Brophy, 1995; Kohn, 1993; Raymond & Hanushek, 2003; Sheldon & Biddle, 1998), teachers (Edelman, 1997; Herman & Golan, 1993; Pedulla, Abrams, Madaus, Russell, Ramos, & Miao, et al., 2003; Smith, 1991), curriculum and instruction (Amrein & Berliner, 2003; Bernauer & Cress, 1997; Madaus, 1988; Vornberg & Hart, 2000), as well as dropout, retention, and graduation rates (Garan, 2004; Goldschmidt & Wang, 1999; Nichols, Glass, & Berliner, 2006, Robelen, 2000; Tienken & Rodriguez, 2010). However, I chose to focus specifically on the impact high-stakes testing has had on student learning and achievement.

Effects and influences on student learning.

Non-experimental empirical studies.

Amrein and Berliner (2002a, b), along with other educational researchers (Braun, 2004; Carnoy & Loeb, 2002; Nichols, Glass & Berliner, 2006; Rosenshine, 2003) aimed to assess how high-stakes testing protocols have influenced and effected student achievement, and more specifically they attempted to answer the question: Are students learning more since the widespread inception of high-stakes testing?

Amrein and Berliner (2002a) differentiated education from merely training students. A clear distinction is established, linking training with acquisition of "useful skills" and associate education with "engagement in cognitive activity that is more demanding than the ability to employ skills" (p.4). Amrein and Berliner (2002a) maintain that high-stakes testing does little more than create a "training effect" on student achievement.

The purpose of Amrein and Berliner's (2002a) non-experimental study was twofold: (a) to evaluate the academic improvements attained (or not) by the 28 states with the highest stakes attached to their grade 1-8 testing policies; and (b) to determine the effect, if any, high school graduation exams has had on improved student achievement. For the purpose of their research, Amrein and Berliner (2002a) defined high-stakes "as consequences that are attached to tests beyond the accountability measures that have been in place for years, like publishing school and district test sores in the newspaper" (p.5).

In an effort to create consistency among the available sources used to establish growth in student achievement since the inception of high-stakes tests in grades 1-8, Amrein and Berliner (2002a) opted to use the National Assessment of Educational Progress (NAEP) results of each applicable state. Although all states are required by NCLB mandates to annually monitor student academic growth through state testing, there is no national level format used. Therefore, using tests that vary state to state to measure academic achievement, comparisons across state lines using data from each state's test would be unethical. Rather than risk analyzing student achievement data that may be flaw or inflated as a result of manipulated curricula, instruction influenced by test preparations, and the accurate representation of special education and limited English proficient learners, an independent measure, although not immune to technical flaws, NAEP was used to conduct the analysis.

Amrein and Berliner (2002a) found that due to rewards and sanctions for school personnel and students attached to the academic performance measured by the state mandated tests, it was inevitable that state scores demonstrated an increase in academic achievement and therefore lacked authenticity. Scores from the ACT, SAT, and Advanced Placement program were assessed at the high school level to investigate the effects of high-stakes testing in high school graduation.

Amrein and Berliner (2002a) included a table that summarized the various possible consequences that are included in testing policies of 28 states that enforce high stakes tests (27 states were actually included in the study, no information was available at the time for the state of Minnesota although in the year 2000 high school graduation became contingent upon an exit exam). A state is classified "high stakes" according to the severity of consequences attached to student performance on state mandated tests (Amrein & Berliner, 2002a). When academic goals are not met and low test scores prevail, sanctions may be bestowed upon the school, faculty and/or the students. A lack of academic improvement may reflect the following: having the state take over, close, or revoke a school's accreditation, or reconstitute low-scoring schools, the replacement of administrators and teachers, grade retention for students, as well as student transfer privileges. Rewards, although less common, are also awarded for high performing schools, staff, and students. In 16 of the 28 analyzed states, monetary awards were given to schools that improved or were high performing; in 8 of the 28 states monetary awards were may be used for teacher bonuses; and 6 out of the 28 states award high performing students scholarships for college tuition. There was no mention of a control group in the researchers' study, which raises concerns in method and design and results in other researchers challenging their findings.

Amrein and Berliner (2002a) consider the fact that policies on excluding students with disabilities and limited English proficiency learners from participating on the NAEP vary by state, which lead them to classify states as unclear if "changes in scores after the introduction of high-stakes tests are related to the rates by which students are exempt or participate in these tests" (p.15). If changes in scores are unrelated to the exemption or participation rates "the effect of high-stakes tests are classified as increases or decreases and weak or strong" (p.15).

In an effort to clarify the approach Amrein and Berliner (2002a) used to conduct an analysis of the data, Braun (2004) summarized by illustrating an example:

> Using NAEP mathematics results for grade 4, they compute the change for the nation, and for each state, over the period 1992 to 2000. They then calculate the differential gain for each state as: State Gain = (change for state '92 to '00) – (change for nation '92 to '00). A positive State Gain means that over this time period the state's improvement on NEAP exceeded that of the nation. Conversely, a negative value means that the nation's improvement exceeded that of the state. (p.4)

Amrein and Berliner (2002a) concluded that "after the implementation of high-stakes tests, nothing much happens" (p.57). According to results collected, an inconsistent pattern emerged among those states deemed "high-stakes," regarding increases and decreases in achievement for math and reading. Amrein and Berliner's (2002a) study refutes the preconceived notions of policymakers, providing evidence that attaching penalties and rewards to student performance on high-stakes tests does not lead to increased academic achievement in schools across the United States. Amrein and Berliner (2002a) identify the need for policymakers, researchers, and education leaders to develop an effective approach to educational reform other than attaching even more stakes to an already flawed system of measurement and accountability.

In an attempt to expose a methodological flaw in the 2002 work of Amrein and Berliner, Rosenshine (2003) reanalyzed the data used in their study. Rosenshine contended that Amrein and Berliner (2002a) did not account for a control group in their study and therefore their findings regarding the impact of high-stakes testing on academic achievement may not be accurate. In the analysis conducted by Rosenshine, unlike Amrein and Berliner (2002a), he comprised states that did not attach consequences to high-stakes tests into a comparison group. Contrary to the findings of Amrein and Berliner (2002a), Rosenshine claimed his analysis revealed "that states that attached consequences outperformed the comparison group of states on each of the three NAEP tests for the last four-year period" (p.1). Therefore, indicating a link exists between high-stakes testing consequences and increased academic achievement.

Using the same NAEP data, Rosenshine (2003) focused on the mathematic gains from cohort to cohort for the years between 1996 and 2000 and used cohort tracking for the years between 1994 and 1998 for the reading test. Rosenshine concluded that "the average NAEP increases in the 'clear' high-stakes states were much higher than the increases in the comparison states" (p.2). Rosenshine reported moderate to large effect sizes including .61 for grade 4 reading, .35 for grade 4 mathematics, and .79 for grade 8 mathematics. In an effort to debunk the reported "decreases" in academic achievement after the implementation of consequences in high-stakes states according to Amrein and Berliner (2002a), Rosenshine considers the ambiguity of the term "decrease". Case in point. Amrein and Berliner (2002a) affirmed that "grade 4 math achievement decreased" (p.36) in Nevada in their study, when in fact grade 4 NAEP mathematic scores increased three points between the 1996 and 2000 period (Rosenshine, 2003). At the time the national average increase was four points, therefore with only a three point increase, Amrein and Berliner (2002a) placed Nevada on the list of decreased math achievement for grade 4 (Rosenshine, 2003). Rosenshine determined similar findings regarding the state of Alabama and grade 4 reading achievement increases.

Rosenshine (2003) reported that Amrein and Berliner (2002a) discovered a total of eight states that decreased in academic achievement in grade 4 mathematics. Rosenshine counters Amrein and Berliner's (2002a) claims by exposing a nonexistent decrease in any of the those states emphasizing that between 1996 and 2000 one state reported a flat change in scores whereas seven of the states classified as incurring "decreases" actually displayed one to four point increases during the 1996-2000 time frame (Rosenshine, 2003).

Rosenshine (2003) maintains that "although attaching accountability to statewide tests worked well in some high-stakes states, it was not an effective policy in all states" (p. 4). Rosenshine's reanalysis added limited evidence to the relationship between high-stakes testing and student achievement. It appears as if Rosenshine's research has merely uncovered even more inconsistencies prevalent amongst states rather than disclose if high-stakes testing leads to increased student achievement. Although Rosenshine acknowledged that the results from his study lack consistencies from state to state, he did discredit the Amrein and Berliner (2002a) accusation that "students are learning the content of the state-administered tests and perhaps little else" (p.58) in states that did display an increase in academic achievement.

Rosenshine (2003) suggested that perhaps the increase in academic achievement witnessed in some of the high-stakes states does not have anything to do with test preparation, consequences, or accountability, but is the result of the return to the strong "academically-focused classrooms" (p. 4), in conjunction with statewide and district polices.

In a rebuttal to Rosenshine's (2003) study, Amrein-Beardsley and Berliner (2003) reexamined the same NAEP data as used in their 2002 study, however this time employing the control group included in Rosenshine's analysis. Amrein-Beardsley and Berliner defended their original position that "high-stakes tests do not do much to improve academic achievement" (p.1). Amrein-Beardsley and Berliner concede that by using states without high-stakes tests as the control group rather than the national trend line as they previously did, Rosenshine was better able to analyze the data and therefore they followed suit with their reanalysis. Amrein-Beardsley and Berliner accepted Rosenshine's findings, and emphasize that he was not incorrect in his calculations, but merely that he worked with the information he had access to at the time.

Re-running their analysis using the control group parameters setup by Rosenshine (2003), Amrein-Beardsley and Berliner (2003) recalled their original concern regarding the exclusion of students sampled to partake in the NAEP tests. Although they concluded that on the grade 4 test high-stakes states (change in score of +4.3) outperformed the control group (states without high-stakes tests) during the 1994-1998 periods (change in score of +2.1), data results yielded otherwise.

Amrein-Beardsley and Berliner (2003) alleged that "because states with high-stakes tests are those states that increasingly are exempting more students from participating in the NAEP" (p.5) it is imperative to analyze the data of states that yielded "clear" effects. The data included in the study indicated that the gains between high-stakes states and the control group were actually insignificant when accounting for and removing the "unclear" states' results, at states without high-stakes tests at +1.6 and states with high-stakes tests at +.5 change in scores (Amrein-Beardsley & Berliner, 2003).

A look at the NAEP Grade 4 math scores from the 1996-2000 period validate Rosenshine's (2003) findings concluding that states with high-stakes tests are outperforming the states that do not enforce high-stakes testing policies on the NAEP. This holds true even when Amrein-Beardsley and Berliner (2003) adjust for the "clear" and "unclear" states in their analysis when results reveal that states with high-stakes tests had a +4.6 change in score while states without high-stakes tests only experienced a change in score of +1.1. Amrein-Beardsley and Berliner maintained that states are targeting students to prevent them from participating in the NAEP Grade 4 math test, and therefore validates their belief "that states with high-stakes tests are not all gaining in NAEP scores simply because of their high-stakes testing policies" (p. 9).

When Amrein-Beardsley and Berliner (2003) re-ran their analyses using the Grade 8 math scores, results disclosed a change in score of high-stakes testing states of +5.4 and a change in score of +1.2 for states without high-stakes tests. Once again, Amrein-Beardsley and Berliner emphasized that the rate of exclusion for particular students must be accounted for to accurately determine the effect high-stakes testing policies had on increased student achievement. Using the same "clear" and "unclear" criteria, an analysis of the states coded as clear/unclear yielded a change of score of +3.0 (states with high-stakes tests) and +.7 (states without high-stakes tests), however Amrein-Beardsley and Berliner reported that the outperformance is not at a "statistically significant level" for grade 8 math (p.12).

Amrein-Beardsley and Berliner (2003) were steadfast in their belief that high-stakes testing states continued to exempt more students from taking the Grade 8 math NAEP test. In fact, they reported "thirty-three percent of the states without high-stakes tests exempted more students and realized gains in math grade 8 NAEP scores. Fifty percent of the states with high-stakes tests exempted more students and realized gains in math grade 8 NAEP scores" (p.12).

After an investigation using the NAEP Grade 4 reading scores from 1994-1998, grade 4 math scores from 1996-2000, and grade 8 math scores from 1996-2000, Amrein-Beardsley and Berliner's (2003) findings reveal "states with high-stakes tests seem to have outperformed states without high-stakes tests in the grade 4 math NAEP at a statistically significant level" (p.12). However, in reference to the grade 4 reading and grade 8 math tests there is no indication that high-stakes states are outperforming states without high-stakes testing policies (Amrein-Beardsley & Berliner, 2003). After the reanalysis of NAEP scores and as a response to Rosenshine (2003), Amrein-Beardsley and Berliner remained "unconvinced that the high-stakes tests used by states are showing systematic positive affects on audit tests used to assess transfer" (p.12).

Braun (2004) is yet another researcher to challenge the findings of Amrein and Berliner (2002b) regarding the impact of high-stakes testing on student academic achievement. When reviewing their work for his reanalysis, Braun used Amrein and Berliner's (2002a, 2002b) classification system, discussed in detail on the previously cited studies, when identifying high-stakes testing states.

Braun (2004) recalled the results Amrein and Berliner (2002b) reported in the study revealed that there were eight states that projected positive state gains, three states that demonstrated negative and two that were reported as zeroes. Braun emphasized that Amrein and Berliner (2002b) described the data from five states as "not available"; however he questioned this finding considering two of these states (Indiana and Minnesota) did in fact have data available regarding the NAEP. When conducting his reanalysis of Amrein and Berliner's (2002b) study, Braun included the data from Indiana and Minnesota that was omitted by the original researchers.

Braun (2004) focused his reanalysis on the math NAEP scores and did not include information regarding the NAEP reading assessment in his study. Braun also conducted a separate analysis following cohorts of students from 1992 Grade 4 mathematics and 1996 Grade 8 mathematics, as well as 1996 Grade 4 mathematics and 2000 Grade 8 mathematics. Braun approached the analysis differently than Amrein and Berliner (2002b) in which his:

> Interpretation of the State Gains statistics is informed by consideration of the corresponding estimated standard errors. (Since the State Gain is a 'difference of differences,' these standard errors are not negligible, with a typical value of 2.5 points on the NAEP scale). (p.5)

By examining the data when computing the state gain and its estimated standard error, Braun concluded that high-stakes testing states that participated in the 1992 and 2000 NAEP mathematics assessment "typically showed improvement relative to the nation while low-stakes testing states that participated in the NAEP mathematics assessment in both 1992 and 2000 typically showed lack of improvement relative to the nation" (p.8). Although similar findings were reported by Amrein and Berliner (2002a, 2002b), these results were discredited based on the exclusion of special education and limited English proficiency learners. Braun acknowledged the concerns raised by the exclusion percentages mentioned by Amrein and Berliner (2002b) however, he argued that there may be other reasons aside from exclusion rates that may account for growth in student achievement that must be considered, such as differences that exists among states that are not observable.

Braun (2004) asserted that for Grade 8 mathematics there is a greater association between gains and high-stakes testing states than in Grade 4, however when tracking cohorts he found that "high-stakes testing effects largely disappeared" (Nichols, Glass & Berliner, 2006, p. 6). Although he presented conflicting data, Braun maintained "with the data available, there is no basis for rejecting the inference that the introduction of high-stakes testing for accountability is associated with gains in NAEP mathematics achievement through the 1990s" (p. 29).

In an effort to examine the relationship between high-stakes testing policies regarding school sanctions and rewards tied to assessment results and student achievement gains, Carnoy and Loeb (2002) developed a "zero-to-five index" used to measure each state's accountability "strength". In the previous studies, reanalysis were conducted using the same classification system as Amrein and Berliner (2002a, 2002b), rather than continue that practice, Carnoy and Loeb's "0-5 scale captures degrees of state external pressures on schools to improve student achievement according to state-defined performance criteria" (p.311).

A vague, general description is provided for each index value, with which Nichols, Glass, and Berliner (2006) find fault. Nichols et al. noted that "Carnoy and Loeb provided very limited information on to how they differentiated a 5 score from a 4 score and so on" (p.7). Nichols et al. (2006) continued their criticism of Carnoy and Loeb's (2002) index system emphasizing that "their index, as a measure of existing laws, did not account for law enforcement or implementation" (p.7).

Using the same data from the 1996-2000 NAEP mathematic assessments as the previously discussed studies, Carnoy and Loeb (2002) conducted a series of regression analyses and determined that gains in mathematical achievement, specifically for eighth grade African American and Hispanic students, were significantly related to strength in accountability. Carnoy and Loeb reported:

> For African Americans the potential gains on the 8th grade test from increased outcome-based accountability are approximately five percentage points for every two step increase in accountability, relative to an average gain of 5.7 and a standard deviation of 5.3. For a two-step increase in the accountability index, the

> gain for Hispanic 8th graders is almost nine percentage points. The mean of the gains is 6.1 percentage points, and the standard deviation of gains among states is 8.5 points, so a two-step increase again makes a large difference. (p. 313)

Sharon L. Nichols, with the University of Texas at San Antonio, and Gene V. Glass and David C. Berliner (2006), with Arizona State University, also conducted a study on the impact high-stakes testing had on student achievement, adding to the limited empirical research currently available in the field. Nichols et al. developed their own accountability pressure rating that was used to rank all 25 states included in their study "based on a continuum of 'pressure' associated with the practice of high-stakes testing" (p.10).

Nichols et al. (2006) devised informative portfolios for each state included in the study that detailed recent and previous practices concerning assessment and accountability. An essay summarizing each state's assessment and accountability plan, a worksheet identifying rewards and sanctions, and newspaper articles were the three sections the portfolios were comprised of. Once the accountability pressure rating system was implemented, Nichols et al. conducted a series of regression and correlation analyses to obtain what relationship, if any exists between high-stakes testing policies and increased student achievement on the NAEP grades four and eight reading and mathematics assessments. Four of Nichols et al.'s key findings were:

1. States with greater proportions of minority students tend to implement accountability systems that exert greater pressure.
2. Increased testing pressure is related to increased retention and dropout rates.

3. NAEP reading scores at the fourth-and eighth-grade levels were not improved as a result of increased testing pressure.
4. Weak correlations between pressure and NAEP performance for fourth-grade mathematics and the unclear relationship for eighth-grade mathematics are unlikely to be linked to increased testing pressure. (Association for Career & Technical Education, 2006, p.9)

The research conducted by Braun (2004), Carnoy and Loeb (2002), and Nichols et al. (2006) all uncovered a link between high-stakes testing and mathematic achievement, "especially among fourth graders and particularly as accountability policies were enacted and enforced in the latter part of the 1990s and early 2000s" (p.51). Nichols et al. provided a possible reason as to why this may be the case; because the mathematics taught at the fourth grade level are skills that could easily be obtained and can demonstrate increased student achievement by repeated drills and practice activities mimicking the actual test.

Raymond and Hanushek (2003) greatly criticized the results of Amrein and Berliner's (2002a) study that were broadcasted across the country via the *New York Times* 2002 headline which warned high-stakes testing "does little to improve achievement and may actually worsen academic performance…" (p.48). In an effort to shed light on what Raymond and Hanushek considered "deeply flawed research" (p.48) they claimed to have uncovered that Amrein and Berliner's (2002a) study "used scientifically inappropriate methods"(p.53), including the absence of a true comparison group, the omission of a blind peer review panel, and cohort tracking inconsistencies.

Raymond and Hanushek (2003) affirmed that by simply using the same approach as Amrein and Berliner (2002a) did, but using all of the available NAEP data "correctly" will reverse Amrein and Berliner's (2002a) conclusions, as a result a "brighter picture" emerges regarding high-stakes testing and increased student achievement. Raymond and Hanushek continuously claimed that Amrein and Berliner (2002a) "ignored" or "overlooked" data they deemed valuable. By recreating an analysis using that questionable data, Raymond and Hanushek presented their findings in an article published in *Education Next*. However, the work of Raymond and Hanushek published in *Education Next* must be cautiously regarded considering it is published by a think tank that has its own education agenda.

Similar to what Rosenshine (2003) and Amrein-Beardsley and Berliner (2003) did in their studies, Raymond and Hanushek (2003) tracked the same cohort of students when analyzing the math achievement gains using the Grade 4 results of 1996 compared with the Grade 8 results of 2000. It is important to mention that the students tested in 1996 do not necessarily make up the sample population of the students tested in 2000; therefore it is not a true cohort tracking. Raymond and Hanushek concluded that significance testing (which they contend is a basic tool used for social-science research), which was not conducted by Amrein and Berliner (2002a), proved that states with high-stakes testing policies displayed larger mathematics gains than no-accountability states in both the 1992-2000 and 1996-2000 time frames. Raymond and Hanushek also took in account the fact that Amrein and Berliner (2002a) might have found fault in their work due to exclusion rates, therefore they provided results that adjusted for changes in students excluded from NAEP testing and reported minuscule changes. At a .05 statistically significant level, in Grade 4 mathematics from 1992-2000, states with high-stakes testing had a 5.3 point advantage over no-accountability states; for the 1996-2000 period there was a 1.9 high-stakes advantage. Grade 8 1992-2000 results revealed a 4.8 point high-stakes advantage and a 2.8 advantage during the 1996-2000 time frame. Minor decreases occurred when adjustments were made to account for exclusion rates.

Raymond and Hanushek (2003) continued to defend the academic improvements that high-stakes testing states incurred by questioning the methods Amrein and Berliner (2002a) employed to make valid before-and-after state comparisons based on high-stakes testing inception dates. Raymond and Hanushek suggested all five states Amrein and Berliner (2002a) labeled suffered "decreases" in student achievement inferring harm was caused by implementation of high-stakes testing, actually exceed gains made by no-accountability states, refuting Amrein and Berliner's (2002a) findings.

Raymond and Hanushek (2003) reminded readers to be cautious of what the media prints regarding such a controversial topic, however the fault lay with Amrein and Berliner (2002a) for neglecting to release a study of sound scientific quality. Raymond and Hanushek found the problem not to be with high-stakes testing, but with how the data from those high-stakes tests are used, reported, and analyzed, asserting "the evidence points in the direction of refining accountability systems rather than scrapping them altogether" (p.55). It is important to remain cognizant of researchers' intentions prior to accepting their findings, suggestions, and criticisms. Raymond and Hanushek warn readers to cautiously read what is printed, however it must be acknowledged that they conducted "questionable research" for a right wing think tank advocacy group and purported opinions which were not grounded on empirically sound research.

Synthesis.

The aforementioned studies all attempted to uncover and divulge information on how student achievement has increased or decreased since the implementation of high-stakes testing mandates. The research currently available concerning the impact high-stakes testing has had on student achievement is mixed, at best (Nichols, Glass & Berliner, 2006). Studies have been conducted that demonstrate no evidence that a relationship exists between states that enforce high-stakes testing policies and increased student achievement gains (Amrein & Berliner, 2002a,b). Yet other researchers believe that there is not enough evidence to discredit the findings that the implementation of high-stakes testing policies positively or negatively impact student achievement (Braun, 2004). There are other researchers that have found their results to indicate that the effects of high-stakes testing may vary according to ethnicity, for example Carnoy and Loeb (2002) reported that high-stakes testing appeared to positively impact achievement in African American and Hispanic students after accountability measures were implemented in certain states. While other researchers steadfastly believe "rigorous analysis reveals that accountability policies have had a positive impact on test scores during the past decade" (Raymond & Hanushek, 2003, p. 50).

Raymond and Hanushek (2003) clearly expressed their disagreement with Amrein and Berliner's (2002a) research that exposed students' academic growth was either unaffected or at times harmed by the implementation of high-stakes testing. Raymond and Hanushek remained adamant in their belief that attaching high-stakes to tests increased student academic achievement and aimed to prove it in their analyses of the 1992-2000 NAEP mathematics student result data.

When considering NAEP student score results, according to the National Center for Education Statistics (2010) the: (a) average writing scale score results are based on the NAEP writing scale, which ranges from 0 to 300; (b) average reading scale score results are based on the NAEP reading scale which ranges from 0 to 500; and (c) average mathematics scale scores are based on the NAEP mathematics scale which ranges from 0 to 500. Recall that Raymond and Hanushek (2003) reported in their study that NAEP mathematics scores revealed a 4.8 point high-stakes advantage in Grade 8 from the 1992-2000 time frame and a 2.8 point advantage during the 1996-2000 time frame. In Grade 4 from 1992-2000 Raymond and Hanushek reported a 5.3 point advantage and a 1.9 point for the 1996-2000 span. Although Raymond and Hanushek attempted to disprove Amrein and Berliner's (2002a) theories of negative high-stakes influences on student learning, what they did instead was emphasize just how insignificant the point advantages truly were. Acknowledging the NAEP mathematics scores range on a 0 to 500 scale, even at the greatest recorded point advantage of 5.3 for the Grade 4 test during the 1992-2000 period, that is still only a 1.06% increase. Therefore, this "advantage" proves completely insignificant to use as evidence of a positive correlation between increased student achievement and the implementation of high-stakes attachment to tests like Raymond and Hanushek affirm. It is alarming that despite these meager findings, Raymond and Hanushek maintained "rigorous analysis reveals that accountability policies have had a positive impact on test scores during the past decade" (2003, p.50).

The adoption of various research designs selected by researchers may be the cause of the mixed conclusions surrounding what relationship, if any exists between high-stakes testing policies and increased student achievement (Nichols, Glass & Berliner, 2006).

One consistency among the researchers is the call for additional empirical studies to assist in determining the impact high-stakes testing policies truly has upon student achievement (Amrein & Berliner 2002a,b; Braun, 2004; Nichols, Glass & Berliner, 2006; Rosenshine, 2003). In this review there were at least six different methods used to investigate the impact that high-stakes testing policies on student learning. Just like all the researchers opted to use the NAEP results to explore this relationship, perhaps developing a unified, agreed upon method for analyzing and reporting data would alleviate many of the conflicting reports and accurately inform the public of how high-stakes testing has impacted student achievement thus far.

As more pressure is placed upon obtaining one hundred percent proficiency for all students, the positive or negative consequences of high-stakes testing policies, whether intentional or not, need to be researched immediately. There is a growing body of research that suggests that high-stakes testing is highly detrimental to the future learning of American minority students and those coming from low-SES families (Amrein & Berliner, 2002ab; Coleman et al., 1966; Darling-Hammond, 2004; Hong & Youngs, 2008; Marchant, 2004; Michel, 2008; Paulson & Marchant, 2009; Popham, 1999; Powers, 2004; Shepard, 2000; Tienken, 2008a; Tienken & Rodriguez, 2010). However, with no end in sight for the taking of high-stakes tests, that does not mean that the conversation has to stop among educators, researchers, policymakers, and community members debating the ethical usage of these tests for high-stakes decision making. High-stakes tests may not be going anywhere anytime soon, however the future academic careers of American students does not have to follow suit. We must contemplate at what point do reported minuscule gains outweigh the negative ramifications associated with attaching high-stakes to standardized assessments?

Due largely in part to the No Child Left Behind (NCLB) Act regulations, and the need to measure student academic growth annually, the sole indicator of success unfortunately, has become the percentage of proficient students on state standardized assessments district and statewide (Paulson & Marchant, 2009). Regarding accountability protocol, Paulson and Marchant (2009) recounted how standardized testing "has been heralded as *the* universal tool" (p.3) for measuring this. However, these high-stakes tests only assess small parts of the curricula, those that can be easily quantitative and broken up into component parts. High-stakes tests do not measure and assess ethics, empathy, character, social consciousness, strategizing, persistence, motivation, collaboration, compassion, or cultural literacy.

Socioeconomic Status and Student Achievement

The landmark study *Equality of Educational Opportunity*, (Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld, & York, 1966) more commonly known as the Coleman Report, issued under the President Lyndon B. Johnson's Administration in 1966 is one of the most cited publications in academic journal articles to date with the number exceeding 2,700 (Gamoran & Long, 2006). In an attempt to uncover what many believed was common knowledge in the late sixties, that poor and minority students were performing badly in school due to a lack of resources, Coleman and his colleagues conducted the large study for the USDOE. Instead the researchers discovered that schools had a small effect on student achievement when other factors such as, student socioeconomic status was taken into account. Coleman and his colleagues reported that the level of success achieved by students on test scores correlated not primarily with school resources and teacher characteristics, but directly with a student's SES and family background.

The NJDOE uses a district grouping system to describe the relative wealth of the community in which each school district is located. The NJDOE recognizes that not every community in New Jersey has the ability to support public education at the same monetary levels. Some towns simply have larger tax bases and ability to pay higher taxes than others. Thus, the District Factor Grouping system (DFG) was introduced by the NJDOE (2006a) in 1975. Analysis of district-to-district test scores and equitably spending provisions are based on the DFG system. The NJDOE identified measurable quantities to create an index to determine a district's DFG: (a) percentage of population with no high school diploma, (b) percent with some college, (c) occupation, (d) population density, (e) income, (f) unemployment, and (g) poverty.

The ranking system includes eight groups: A, B, CD, DE, FG, GH, I, and J; ranging from the lowest socioeconomic status districts to the highest. In a way, the DFG system helps to identify the potential inequities brought on by various degrees of poverty. Those who understand Coleman et al.'s findings can see similar results play out in the test results from New Jersey's statewide tests, as the results fall out along DFG lines (Tienken, 2008a).

Empirical studies.

The 749 page Coleman Report (1966) contained an array of information detailing: school environment (i.e., school facilities, services, curriculum, staff, and fellow students), pupil achievement and motivation (i.e., outcomes of schooling, integration and achievement), future teachers of minority groups, higher education, non-enrollment records, case studies of school integration, and special studies, among other various findings, however the most controversial was the discovery that once SES was controlled for, school resources have very little influence on academic performance (Gamoran & Long, 2006). Coleman et al. (as cited in Gamoran & Long, 2006) conducted an analysis "by measuring the proportions of variance in student achievement that could be attributed to school facilities, school curriculum, teacher qualities, teacher attitudes, and student body characteristics" (p.7). Through questionnaires and surveys and by aggregating data from 60,000 teachers and 570,000 students, Colman et al. (as cited in Michel, 2004) found that:

> Socioeconomic status explained a greater proportion of student test scores than other measures of school resources such as class size and teacher characteristics; 49% student background, approximately 42% teacher quality, and 8% class size. The report showed that a school's average student characteristics, such as poverty and attitudes toward school often had a greater impact on student achievement than teachers and schools, and that the average teacher characteristics at a school had a small impact on a school's mean achievement. (p.29)

Thirty-six years after the Coleman Report, Goldhaber (2002) reported that 60% of the variance in student achievement was directly associated with student SES and family background, followed by 8.5% of the variation due in part to teacher characteristics.

In an effort to gain "a 40-year perspective" on the *Equality of Educational Opportunity,* Gamoran and Long (2006) documented information regarding: (a) how Coleman's (1966) original findings hold up 40 years later after numerous subsequent research has been conducted; (b) the international value of the Coleman Report; (c) Coleman's debate over school choice and vouchers; and (d) 40 years worth of equality of educational opportunity and contemporary education reform policies. For the purpose of my research, I chose to focus only on the information that pertained to numbers one and four. Coleman et al. (1966) found that 85% of Black students whom received an education through the 12th grade scored below the average for Whites. Gamoran and Long emphasized "on average, Blacks scored a standard deviation below Whites in academic achievement" (p.5). The majority of tests used to measure student achievement are classified as "norm-referenced". Norm-referenced tests compare an individual student's performance with that of others in order to establish meaning of the score values. "The standard deviation is a statistic that provides an overall measurement of how much participants' scores differ from the mean score of their group" (Pyrczak, 2006, p.49). Therefore, the difference of one standard deviation in terms of black and white student scores is extremely significant, in fact this equates to the difference between scoring at the 50th percentile and 84th percentile. The researchers maintained that the once highly prevalent achievement gap between White and Black students examined by Coleman et al. have since narrowed. NAEP results indicate that the once predominate gap in reading achievement for 17 year-old Black and White students was 1.2 standard deviations in 1971 but reduced to .69 by 1996 (Gamoran & Long, 2006). Jencks and Phillips (1998) reported that mathematics achievement gaps between Blacks and Whites declined from 1.33 to .89 standard deviation units as well. Gamoran and Long (2006) recalled that these achievement gap declines occurred during the 1970s through the 1980s, ironically those gaps increased during the 1990s. As of 2004, the gap in Black and White NAEP mathematic scores for 13 year-olds consisted of a 27 point difference, whereas a 22 point difference existed for reading scores (Perie, Moran, & Lutkus, 2005).

After the publication of the Coleman Report (1966) other researchers decided to conduct their own analysis to determine if their findings would simulate those of Coleman et al. Through several investigations, Averch, Carroll, Donaldson, Kiesling, and Pincus (1974) discovered inconsistencies when attempting to identify which school resources dominated the influence on student achievement. Averch et al.'s reported mixed results, however they did arrive to a similar conclusion as Coleman et al., that a student's socioeconomic background is the largest contributor to student success and "that there did not seem to be much value to paying a premium for smaller class size or teacher experience or advanced degrees" (Gamoran & Long, 2006, p.7). Gamoran and Long (2006) also highlighted studies that challenged the findings of the Coleman Report in their 40 year retrospective review. Gamoran and Long summarized:

> These critiques have included arguments that Coleman's cross-sectional study could not adequately capture causal effects, that Coleman assumed a linear and additive relation between resources and learning, that cross-sectional measures of reading achievement could not distinguish between learning that occurs at home and learning that occurs at school, and that Coleman's estimation of school effects by measures of percent of variance explained were sensitive to assumptions about causal ordering (Sorensen & Morgan, 2000; Hanushek, 1979; Hanushek & Kain, 1972; Bowles & Levin, 1968). (p.7)

In 1972 Mosteller and Moynihan shared that it was their belief that one of the most significant findings of the Coleman Report (1966) was that there was very little difference between the resources allocated to Black and White schools, therefore claiming that gaps in achievement are the direct result of some other factor. Although Jencks et al. (1972) agreed with Mosteller and Moynihan that there was value in the Coleman Report findings that determined little variance in resources from Black and White schools existed across the United States, Jencks et al. also found significance in other results brought forth by Coleman and his colleagues, such as the academic achievement increase of students with lower socioeconomic background that attended schools with affluent peers. Jencks et al.'s investigation determined that after measures were taken into account for "sampling procedures, information-gathering techniques, and analytic methods" the Coleman Report results "[held] up surprisingly well" (p.70). Gamoran and Long (2006) recounted how Smith (1972), when reviewing the Coleman Report, "focused on regression coefficients instead of percent of explained variance, [and] came to similar conclusions about the lack of effect of school resources once family background is controlled" (pp.7-8). Researching the impact of different causal sequencing of the Coleman Report variables, Hanushek and Kain (1972) like the previous studies, but in a very different context arrived at the same conclusion that school resources have little effect on student achievement to show that funding for schools should be sharply decreased.

When analyzing educational attainment, Jencks et al. (1972) concluded that family background had such a strong effect on student performance noting that until inequalities pertaining to occupational status, education, and parents' income vanish, inequalities will continue to exist in educational institutions. In the 1990s educational researchers continued to analyze the Coleman Report findings for the effects school resources has on student performance (Greenwald, Hedges, & Laine, 1996a, 1996b; Hanushek, 1994, 1996, 1997). Greenwald, Hedges, and Laine (1996a) criticized the findings of Hanushek's (1981, 1986, 1989, 1991) multiple publications that claimed there was no relationship between school resources and student achievement. Greenwald et al. expressed their dissatisfaction with the fact that Hanushek's highly flawed synthesis method actually gained substantial recognition and acceptance by some in the legal, academic, and public policy forums.

Greenwald et al. (1996a) explained that "Hanushek's synthesis method, vote counting, consists of categorizing, by significance and direction, the relations between school resource inputs and student outcomes (including but not limited to achievement)" (p.362). Greenwald et al. criticized Hanushek's "vote counting" method identifying it as an outdated, "rather insensitive procedure for summarizing results" (p.362, as cited in Hedges & Olkin, 1980). After conducting a reanalysis of Hanushek's (1986) conclusions, Greenwald et al. affirmed "that the data on the relations between school resource inputs and student outcomes, including achievement, were substantially more consistent and positive than he believed" (p.362). How then, using similar data sources, did these researchers arrive at conflicting conclusions? Greenwald et al. also addressed that in their meta-analysis.

When preparing the inclusion criteria for their meta-analysis, Greenwald et al. (1996a) considered the following:

1. The data are presented in a refereed journal or a book.
2. The data originate in schools in the United States.
3. The outcome measure is some form of academic achievement.

4. The level of aggregation is at the level of school districts or smaller units.

5. The model controls for socioeconomic characteristics or is either longitudinal (including a pretest and a posttest) or quasi-longitudinal (including IQ or a measure of earlier achievement as an input).

6. The data are stochastically independent of other data included in the universe. (pp. 364-365)

Out of the 38 studies employed by Hanushek (1986), 29 studies were used in Greenwald et al. (1996a) meta-analysis, even though "many equations and coefficients failed to satisfy the decision rules" (p.363). Greenwald et al. concluded that their meta-analysis confirmed "that school resources [were] systematically related to student achievement and that these relations [were] large enough to be educationally important" (p. 394). Greenwald et al. emphasized that although they anticipated their findings would validate a relationship between resources and student achievement they were surprised that the "conclusions [were] so uniform in direction and comparable in magnitude" (p.385).

Greenwald et al. (1996a) also exposed their findings, that school resources were associated with student academic growth, by analyzing NAEP scores. At the time of their study the most recent available NAEP data was from 1992. NAEP trend data has been available in reading from 1971 and in mathematics from 1973. Greenwald et al. recalled that from the early 1970s to 1992 "the national average achievement of White students has remained fairly stable, the national average reading and math achievement of Blacks and Hispanics has increased by about one half a standard deviation" (p.383).

Greenwald et al. emphasized that the increase in student achievement of Black and Hispanic students was "substantial" and that this data coincided with their findings that school resources positively influence student achievement.

In a response to Greenwald et al. (1996a), Hanushek (1996) stated that the results they presented were "distorted and misleading" (p.397) and did not validate their belief that school resources positively influenced student performance; rather it reinforced his original findings. Hanushek lamented his usual concerns and displeasure at how Greenwald et al.'s work has added little to the field of educational research, except to increase the level of confusion throughout the nation. Hanushek stressed from the onset that Greenwald et al.'s meta-analysis suffered methodological flaws that ultimately lead them to mistakenly believe:

> (a) that U.S. schools have been working quite well, (b) that schools have been providing a good return on expenditure, (c) that any performance problems of students are best attributed to poorer students and parents and not to the schools, and (d) implicitly, that more resources devoted to the current schools would be productive and would be a wise investment for society to make. (p.398)

Hanushek (1996) asserted that Greenwald et al.'s (1996a) analysis suffered from three major flaws including: (a) a misinterpretation of what their findings imply in the field, (b) a deliberate bias of results in order to achieve their desired conclusions and, (c) "a flawed statistical approach for investigating issues of how and when resources affect student performance" (p.398). Hanushek personally and professionally chastises Greenwald et al. concerning multiple aspects of their analysis. Hanushek began by criticizing the purpose of Greenwald et al.'s "completely uninteresting question" that lead their analysis claiming it had no relevant policy perspective. From there, Hanushek continued by criticizing Greenwald et al.'s sample selection describing how:

> For purely technical reasons their methodology requires that they eliminate all studies finding statistically insignificant effects but not reporting the sign. This action by itself eliminates 13% to 26% of the available data. Clearly, since they are out to show that there is a statistically relationship, the preliminary elimination of substantial evidence to the contrary biases the results in favor of their perspective. (p.400)

Another contention Hanushek (1996) expressed regarding Greenwald et al.'s (1996a) sample selection was how they were "dramatically biased toward retaining both statistically significant positive and insignificant but positive results" (p.402), ironically once again, favoring the conclusions they needed to make their point. In addition, Hanushek chided Greenwald et al. for reporting that students have gotten worse over the years. Hanushek contended that Greenwald et al. attributed increases in the female work force and the increase in single family households, which according to Hanushek "impl[ied] to them poorer family inputs to kids' education" (p.405). What Greenwald et al. overlooked, according to Hanushek, was the fact that the country did witness a dramatic decrease in family size and an increase in parents with a higher level of education than previous years.

Hanushek (1996) continued to condemn Greenwald et al. (1996a) on other portions of their study, such as the inclusion of longitudinal studies, policy issues, and specific meta-analytic methodology. Hanushek summarized his critique of Greenwald et al. by repeating that when the proper, appropriate methods are employed the results still point to "the lack of a consistent relationship" (p.406) between school resources and improved student achievement.

In a rejoinder to Hanushek (1996), Greenwald et al. (1996b) is emphatic that they "do not endorse: that schools are currently working well, that they are providing a good return on investment in education, that performance problems are attributed to poorer students, and that investing more money in current schools would be wise" (p.411). Greenwald et al. meticulously debated the serious charges Hanushek raised regarding their meta-analysis methodology as well as their conclusions.

Analysis issues were one area that the researchers differed greatly. Citing several experts, Greenwald et al. (1996b) defended their use of meta-analytic methods when reviewing production function studies, contrary to what Hanushek claimed. Greenwald et al. recalled that since their 1994 work, Hanushek has since changed his position. At one point Hanushek denied that there was any systematic effect of school resources on student academic performance, then shifted "to agree[ing] that there is distribution of results, with some, perhaps most, of the studies finding a preponderance of schools in which greater resources are associated with greater achievement" (p.419).

On the topic of sample selection, Greenwald et al. (1996b) responded to the accusations that they used what Hanushek (1996) referred to as "a very selective sampling of available results" (as cited in Greenwald et al., 1996b, p.413). Greenwald et al. emphasized that their study clearly reported the "criteria for choosing coefficients, but Hanushek [did] not" (p.413). Taking that into account, Greenwald et al. contended that speculation will continue to shroud Hanushek's work "until he reveals something about the procedures used to obtain publications and extract information from them" (p.413).

As a result, it appeared as if Hanushek had included a greater number of studies in his analysis than Greenwald et al. had considered, however that assumption is false. By counting some results of data sets multiple times, particularly those containing negative results as opposed to those with positive results, "Hanushek is able to achieve the appearance that the evidence is more evenly divided than we found it to be" (p.414).

Greenwald et al. (1996b) found it odd that Hanushek decided to criticize them for not distinguishing between longitudinal and quasi-longitudinal studies, and for neglecting to provide a full description of the selected studies. However, Hanushek's criticisms were unfounded. In fact, descriptions of all the included studies were provided in Appendix A of Greenwald et al.'s work as well as Table 6 that separated and provided detailed information regarding the effect sizes of both longitudinal and quasi-longitudinal studies. Ironically, Hanushek who inaccurately criticized Greenwald et al. neglected to ever mention, yet alone describe, the studies he used to gather the reported results of his meta-analysis.

In their meta-analyses both Greenwald et al. (1996a, b) and Hanushek (1996) arrived at conflicting findings regarding the impact school resources had on student achievement. Greenwald et al. concluded there were correlations between resources and student achievement, whereas Hanushek determined an inconsistent, random pattern regarding the effect of the same variables on student achievement. Gamoran and Long (2006) explained these conflicting reports are the result of the researchers difference in inclusion criteria when selecting studies for their analyses; Greenwald et al. was more selective, whereas Hanushek classified findings of previous studies as negative, positive, or neutral. Greenwald et al. and Hanushek may have differed in some aspects of their findings, however these researchers did agree that:

> (a) In at least some cases, higher levels of resources are associated with higher achievement; (b) the qualities of schools that produce these effects are hard to pin down; and (c) the ways in which resources are used is more consequential for achievement than the presence or absence of resources. (Gamoran & Long, 2006, p.8)

Greenwald et al. (1996a, b) were not the only critics of Hanushek's work. Over the past four decades Hanushek's findings in numerous studies have been criticized and scrutinized by others that find fault with his conclusions. For example, Spencer and Wiley (1981) recalled a study in which Hanushek (1979) "misinterpret[ed] the data on which he base[d] his conclusion and draws inappropriate policy implications from them" (p. 43). Spencer and Wiley argued that Hanushek focused on how while national expenditures for public education were steadily increasing, the academic achievement of students was not following suit, but instead was decreasing. Spencer and Wiley blamed a "widespread misunderstanding of determinants of scholastic performance" (p. 44) for the reason why inaccurate educational production models littered the research, which ironically, Hanushek's analysis was founded upon.

Spencer and Wiley (1981) continued to expose the flaws in Hanushek's (1979) analysis by highlighting the issues that arose from assuming educational productivity analysis was identical to techniques of economic productivity analysis. Regarding the methodological concerns surrounding Hanushek's conclusions, Spencer and Wiley summarized:

> That the fundamental problems of model specification, identification of "inputs" and "outputs," and biased data are so severe that drawing any conclusions from the educational productivity analyses—either singly or in synthesis—is hazardous. Furthermore, the regression coefficients cannot be interpreted as marginal productivities. Accordingly, any conclusions drawn from the data for policy purposes would be baseless. (p. 49)

Like Spencer and Wiley (1981), Baker (1991) also expressed contention with Hanushek's (1986) misuse of data in an effort to support biased educational policy conclusions. In an effort to showcase the "wasted" money that was provided to schools according to the then U.S. Secretary of Education William Bennett, Hanushek's work was referenced as the proof that money should stop being "thrown" at schools. Using Hanushek's findings, Bennett claimed that there was no relationship between spending and student achievement. Baker questioned the work of Hanushek since the data he presented yielded 13 positive and three negative significant results as well as 25 positive, 13 negative, and 11 unknown insignificant results. Baker contended "Hanushek did not explain the decision rule that he used to conclude that these numbers indicate that there is no relationship between spending and achievement. He simply presented the numbers and concluded that there is no relationship" (p.629).

Baker (1991) insisted that there were logical and methodological problems with Hanushek's (1986) work. For starters, Baker claimed that both Hanushek and Bennett attempted to answer the question of how spending effected achievement gains, not levels. Baker explained:

> The difference between the two is that gains are what is learned during the school year, while levels are where a student stands at the end of the year. The distinction is critical, because family background affects levels but not gains. Thus differences between the average achievement levels for schools may reflect only differences in family socioeconomic status while masking the relationship between achievement gains and expenditures. (p.629)

Because Hanushek did not distinguish between levels and gains when making his calculations, according to Baker he used the incorrect data to answer his question. Baker also noted that although Hanushek made several references to the fact that much of the material he reviewed contained methodological problems, he just so happened to overlook that omitted from his work was a rational for arriving at the conclusion that there was no evident relationship between student achievement and school expenditures.

Baker (1991) also questioned the method Hanushek (1986) used to review the available literature. At the time, meta-analysis was a quantitative procedure that was used more often than the traditional method of merely counting results, however Hanushek opted not to include a meta-analysis from his review of the literature. Baker contended that this omission was "a major, though not necessarily fatal weakness" (p. 630) of Hanushek's study.

In a reanalysis of Hanushek's (1986) work, Baker (1991) concluded that because Hanushek did not include a decision rule, he "applied different decisions rules to Hanushek's data with the same result: the more money schools spend, the higher their achievement" (p. 630). According to Baker (1991) in an effort to promote an anti-spending policy, Bennett relied on Hanushek's highly flawed conclusions to propel "the

Reagan Administration's efforts to curtail federal spending on education programs" (p. 630).

Other researchers suggested through their analysis that school resources such as class size and teacher quality have a greater effect on African American and students from lower socioeconomic backgrounds and that variation exists among schools (Finn & Achilles, 1999; Summers & Wolfe, 1972). Jencks and Phillips (1998) look to standardized tests as one cause of the Black-White prevailing gap in academic achievement, however they asserted "schools cannot be the main reason for the gap, because it appears before children enter school and persists even when black and white children attend the same schools" (p.3). This notion is corroborated by a study conducted by Betts, Rice and Zau (2003) that maintained "a first important observation is that students, from very early in their educational experiences, appear to exhibit large variations in achievement that are systematically linked to poverty" (p.8). A 2003 Public Policy Institute of California report confirmed Jencks and Phillips's assertions by disclosing "the daunting achievement gaps between students do not appear to be created primarily by the schools as they now exist. Taking everything into account, income, and socioeconomic status still matter, and they matter a great deal" (Betts, Rice & Zau, p.4). Considering a student's socioeconomic background is not within the control of a school, Goldhaber (2002) asserted that "the most important thing a school can do is to provide its students with good teachers" (p.52). This point is emphasized in a more recent study conducted by Michel (2008). Michel identified that after controlling for student and school variables, the strongest predictor of student performance was the percentage of teachers that held master's degrees.

Forty years after the publication of the hallmark *Equality of Educational Opportunity* (1966) study, Gamoran and Long (2006) stated "the findings of the Coleman Report hold up remarkably well, in some ways distressingly so" (p.19). Gamoran and Long consider the future implication of the startling discovery that not much reform has occurred in equality of educational opportunity 40 years after the release of the Coleman Report. The researchers were confident that change in educational equality is possible by either: (a) enacting country-wide policies that benefit disadvantaged students rather than their more advantaged peers; and (b) through policy revisions that focus on the effects of disadvantaged students rather than all students. Gamoran and Long concluded that although research over the last 40 years varied on the strength of the relationship between school resources and student achievement, when working with disadvantaged students the qualifications of teachers, class size, and school resources may have more influence and therefore should not be overlooked entirely.

Using the same data Marchant, Paulson, and Shunk (2006) continued to investigate further by controlling for family income, parent education, ethnicity, and exclusion of special education and limited English proficient learners, characteristics according to the researchers that are associated with NAEP achievement. Marchant, Paulson, and Shunk (2006) acknowledged that Nichols, Glass, and Berliner (2006) and Carnoy and Loeb (2002) conducted similar studies, and that they did not want to replicate their studies, but rather use what they "considered to be more conventional statistical techniques to demonstrate the importance of considering family income and parent education levels of test--takers in comparing groups (i.e., states) on aggregated achievement data" (p.4). Marchant, Paulson, and Shunk concluded that when analyzing reading and science NAEP results, students in high-stakes testing policy states scored slightly lower than those in states without high-stakes testing policies. However, when compared longitudinally, over a 4 year period, mathematics and reading scores revealed that high-stakes testing could account for significantly improved test scores. Perhaps most interestingly, "further analyses showed that most of these relationships (whether in favor of high stakes or non-high stakes states) disappeared once demographic differences were controlled" (p.22).

Marchant, Paulson, and Shunk (2006) affirmed that many of the previously reviewed studies that investigated the relationship between high-stakes testing and student achievement did not control for demographics, therefore leading to the all too common mixed results. Marchant, Paulson, and Shunk ominously advise:

> When up to 70% of the variability among states' aggregated NAEP scores can be predicated by the average demographic characteristics of the states' test-takers—factors outside of the control of educational policies—educators and policy makers should be careful when attributing differences among states' performance to the policies alone. (p.22)

The results Marchant, Paulson, and Shunk (2006) released confirm the significance of considering family income and parent education which "proved especially valuable in predicting variability among testing samples" (p.22).

Synthesis.

Is failure imminent for children of low SES? Coleman et al. reported in 1966 that the greatest influence on student academic performance was SES, followed by teacher characteristics and class size. Over 40 years after the release of the Coleman Report, much of the reviewed literature continues to support the original findings of Coleman et al., even when attempts were made to debunk those findings. As the debate continues regarding specifically what teacher and school resources influence student achievement the greatest, one aspect of the extent research remain consistently clear, SES is the single strongest predictor of student performance. The very tests that are being used to monitor academic achievement progress and the desired narrowing of the minority achievement gaps may be the "weak link" in the educational reform initiative. After reviewing the extensive literature available regarding the potential attainment of educational equality among students it is evident that enacting accountability policies, providing additional funding, using high-stake consequences and the results from those tests as major indicators of student academic success, and providing an increased number of education resources to struggling schools will not, in and of themselves, lead to the successful bridging of existing achievement gaps at the state and national testing level (Lee &Wong, 2004).

Attendance and Student Achievement

A factor that negatively influences student academic achievement is absenteeism. In fact, the effects of being absent frequently can be so detrimental to student learning that they lead to other risk factors that appear later in life (Dryfoos, 1996; Finn, 1993; Gottfried, 2009; Lehr, Sinclar, & Christenson, 2004; Stouthamer-Loeber & Loeber, 1988). There is a significant amount of literature available examining the relationship between attendance and student performance on standardized achievement tests.

Chen and Stevenson (1995) performed a study that examined the mathematic achievement of Asian-American students from a cross-cultural perspective. The researchers found that attendance was an "achievement-related behavior" that influenced student achievement outcomes on exams; the more often a student was absent the more poorly he or she performed on the exam. The focus of Dryfoos' works pertained to at risk adolescents and behaviors associated with being "high risk." Dryfoos (1996) contended that a relationship exists between academic achievement and attendance noting "because poor school achievement has such a strong negative influence on other outcomes, communities with schools that have high failure and dropout rates are also communities with high delinquency…." (p.10).

In order to obtain information with respect to the influence attendance had on student achievement, Caldas (1993) conducted a study using Louisiana public schools. Included in the study were elementary and middle schools from inner-city and non-central locations. Caldas reported that attendance rate was a significant predictor of student achievement on high-stakes assessments. Roby (2004) conducted a similar study investigating educational outcomes of Ohio students in Grades 4, 6, 9, and 12. Roby also found evidence of a statistically significant correlation between attendance and student academic performance. Also using test data from Ohio, Sheldon (2007), affirmed that reading and mathematic test results were highly correlated to student absences.

Synthesis.

The reviewed literature indicated that there is a significant relationship between student academic achievement and attendance. Empirical evidence exists confirming that as a student's absenteeism rate increases, the poorer they perform academically. Thus, attendance is a strong predictor of student achievement on state mandated assessments.

Gender and Student Achievement

Gender is a variable that researchers still consider when analyzing influences of student achievement. There is no definitive cause for student achievement differences among gender. What is documented is that personal, instructional, and environmental variables account for gender discrepancies (Wilkins, Zembylas, & Travers, 2002). Specifically, these factors include an individual's socioeconomic status (Drukker et al., 2009), culture and surroundings (Pajares, 2002), neurological composition (Gurian & Stevens, 2004), testing regulations established by state and federal agencies (Gunzelmann & Connell, 2006), as well as sociological/biological consideration (Salamone, 2003).

Using data from 31 participating countries, Marks (2008) analyzed the 2000 Programme for International Student Assessment Project (OECD, 2001) results to determine how student achievement in reading and mathematics was influenced by gender. Marks concluded that:

> The gender gaps in reading and mathematics are highly correlated and that the magnitude of the gaps reflect the implementation and success or otherwise of policies designed to improve girls' educational outcomes are likely to reduce the gender gaps in mathematics but increase the gender gap in reading. (p. 106)

The underrepresentation of women in the science, technology, mathematics, and engineering fields is a concern that Americans cannot afford to overlook. Although evidence exists that indicates otherwise, stereotypes still prevail that claim females lag behind their male peers in mathematics achievement (Else-Quest, Hyde, & Linn, 2010; Hedges & Nowel, 1995; Hyde, Fennema, & Lamon, 1990; Hyde, Lindberg, Linn, Ellis, & Williams, 2008). In an effort to obtain more information regarding the magnitude of gender differences in mathematics achievement and attitudes, Else-Quest et al. (2010) examined patterns of gender differences cross-nationally. Analyzing 2003 Trends in International Mathematics and Science Study and PISA results, the researchers determined that "on average, male and females differ very little in mathematics achievement, despite more positive math attitudes and affect among males" (p. 125).

Warren W. Willingham, scientist for Educational Testing Service, and President Nancy S. Cole (1997) conducted a 4 year study analyzing gender differences on assessments. Willingham and Cole exposed several myths surrounding the conventional notion that girls generally do well in the liberal arts whereas boys tend to excel in mathematics and the sciences. Willingham and Cole concluded that data revealed that there was essentially no difference between females and males for 74 assessments at the 12th grade level across 15 subject areas. The gender gaps of the 1960s have since narrowed. The researchers debunked the belief that boys outperformed girls in mathematics and science, but did report a minor advantage for girls in writing. The authors asserted that gender differences do not necessarily account for gaps in student achievement, but that individual factors and personality traits can also influence academic achievement of students.

Synthesis.

Although the stereotype is that females outperform males in liberal arts and that males outperform females in mathematics and the sciences, there is little truth to this. Research shows that although there was once a gender gap in mathematics and sciences it has since diminished. There is little to no empirical evidence concerning the gender gap in liberal arts today. One study reviewed did find that data from 1990 confirmed that girls had a slight advantage in writing over boys.

Formative Assessment and Student Achievement

Formative assessment definitions.

The review of literature pertaining to assessment types resulted in conflicting findings. Although there is a myriad of available literature on formative and summative assessments, also referred to as formative and summative evaluations, there is an overwhelming amount of confusion surrounding the exact definition as to what these terms entail. How can one research and supplement the existing literature with empirical data when the subject is not unanimously understood? Dunn and Mulvenon (2009a) addressed this concern by highlighting a host of formative assessment definitions to display the vagueness of the term used by numerous education researchers. At first glace the definitions contributed to the field by Black and Wiliam (1998), Leung and Mohan (2004), Wininger (2005), Popham (2008), Bell and Cowie (2000), Stiggins (2002) as well as a multitude of other well known researchers, appear strikingly similar. Upon closer examination however, it becomes apparent that this lack of continuity is the reason as to why little empirical evidence is available regarding the effectiveness of formative assessments on student achievement.

Historically, many of the definitions of formative and summative assessments stem from Michael Scriven's (1967) work in which the use of summative and formative evaluations in education were analyzed and embraced by educators (Popham, 2008). Scriven's differentiation among the two terms is echoed greatly in the recent works of the previously mentioned researchers and widely accepted in the today's educational classroom forum. Popham (2008) recounted Scriven's work by highlighting that the primary purpose of formative assessments was to modify instruction during the learning process in order to increase student achievement, while summative assessments evaluate the instructional strategies used to attain the aforementioned increase in student achievement. More specifically, "formative assessment," according to Perie, Marion, and Gong "is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes" (2007, p.1). Perie at al. concluded summative assessments evaluate students' final performance at the end of a unit or school year in relationship to the core curriculum content standards. In their opinion, summative assessments do not permit the use of informative feedback and therefore are viewed as "the least flexible of the assessments" (p.1). One defining characteristic about formative assessment is the frequency of administration. Formative assessment should occur frequently enough to allow teachers to monitor and adjust their lessons (Perie, Marion, & Gong, 2007). Therefore, formative assessment is something that is integrated into instruction daily. The intended purposes, audience, and use of the information distinguishes formative, summative, and interim assessments from each other.

Ironically, it is these "least flexible assessments" that state agencies are required to use to measure students' academic achievements in accordance with the NCLB Act accountability systems. Unfortunately, there is great concern that these summative assessment results are being used for purposes for which they were not originally designed. Perie et al. (2007) recognized that educators are cognizant of the imperativeness of using individual student data to drive instruction, however these educators are mistakenly using summative assessments results, albeit unsuccessfully, rather than formative assessments results as a means to achieving this end. Consequently, it is not plausible that summative assessments could effectively improve classroom instruction. This understanding has led administrators and educators to seek formative or ongoing assessment products that can be use throughout the course of the school year to monitor and track students' achievements earlier rather than later.

Traditional formative assessment and academic achievement research.

Due largely in part to the NCLB Act stipulations, students are formatively assessed more frequently in language arts and mathematics in order to ensure annual yearly targets are met (Perie, Marion, & Gong 2007; Plake, 2002; Sloane & Kelly 2003). It is imperative for policymakers, researchers, and education leaders to ponder the following questions prior to the implementation of formative assessment practices: Does administering various forms of formative assessments (i.e. FAT), ultimately lead to improved student achievement? More importantly, how effective, if significant at all, are formative assessments results in predicating increased student achievement on statewide and high-stakes assessments? The formative assessment studies reviewed will attempt to identify the gaps that exist in the knowledge base while highlighting the strengths and weaknesses discovered in the current literature.

The fundamental and widely cited journal article written by Black and Wiliam (1998) showcased eight empirical studies conducted by other education researchers. According to the authors these studies affirmed the positive implications of using a variety of formative assessment procedures in the classroom with strong evidence that support student gains in achievement are yielded. The examined and cited experimental, quasi-experimental, and meta-analysis studies referenced by Black and Wiliam included: Fontana and Fernandes (1994), Whiting, Van Burgh, and Render (1995), Martinez and Martinez (1992), Bergan, Sladeczek, Schwarz, and Smith (1991), Butler (1988), Schunk (1996), White and Frederiksen (1998) as well as Fuchs and Fuchs (1986).

Conclusions drawn by Black and Wiliam (1998) regarding the significant impact of formative assessment on student achievement as well as the works of the aforementioned researchers, is considered highly flawed by others in the field of educational research. Many of the eight studies relied upon by Black and Wiliam have suffered criticisms as recently as March of 2009 (Dunn & Mulvenon, 2009a).

Experimental studies.

In an experimental study conducted by Fontana and Fernandes (1994) 25 qualified primary school teachers (experimental group) in Portugal were taught self-assessment techniques by one of the authors, Margarida Fernandes. Teachers then employed the learned techniques with "non-selective third and fourth year (ages 8 to 14) primary school pupils over a specified period" (p.408) to determine if students' mathematical academic achievement was influenced by the self-assessment strategies. The control group consisted of 20 primary school teachers that engaged in an alternate curriculum development professional development program, a completely different program than that experienced by the experimental group, but led by the same instructor, Margarida Fernandes.

Fontana and Fernandes (1994) boasted that their findings clearly indicated that those students whom were frequently encouraged to participate in self-assessment strategies "show[ed] a significant advance in academic (mathematics) performance when compared with a control group who d[id] not operate these strategies" (p. 414). Black and Wiliam (1998) reported that the results of the Fontana and Fernandes study demonstrated gains in student achievement double for the experimental group of younger children only that used self-assessment daily as compared to those that did not participate at all in the strategy. Black and Wiliam believed that these results were significant and therefore proved that formative assessment use improves student achievement. However, when I conducted calculations using data provided from the Fontana and Fernandes study and the formula for determining the effect size of an intervention (the mean of the experimental posttest results minus mean of the control posttest results divided by the standard deviation of the control), a different interpretation emerged. The effect size for Grade 3 was -0.03, Grade 4 was 0.14, and the overall effect size of the intervention used in the study was 0.04. Although Black and Wiliam and Fontana and Fernandes purported that the results of this study indicated student achievement was positively influenced by the intervention of formative self-assessment. Once calculations uncovered what was actually a nominal effect size, the findings generated additional, noteworthy questions regarding the usefulness of this particular formative assessment.

However, Black and Wiliam (1998) failed to acknowledge that arriving at "conclusive decisions about the effectiveness of all formative evaluation based" (Dunn & Mulvenon, 2009a, p. 6) on a small sample population and single content area is problematic. An inadequate sample size was not the only problem of using this study as solid evidence for formative assessment practice. The original authors, Fontana and Fernandes (1994), as well as Black and Wiliam recognized that another factor may have influenced the outcomes, such as a too simplistic pretest for the older group of students.

Although the study attempted to correlate the statistical significance of self-assessment techniques and increased student achievement, insufficient evidence exists indicating the impact these measures would have on a larger sample of teachers and subsequent students. All authors failed to mention the effect the different professional development programs may have had on the control group represented in this study (Dunn & Mulvenon, 2009a). It might be less than accurate to make assumptions that the formative assessment strategy of self-assessment, by itself, positively influenced and increased student mathematical academic achievement based on this study, as Black and Wiliam have done, without recognizing that rival explanations exist.

The work of Martinez and Martinez (1992) also contributed to the limited research available on the positive effect of formative assessment usage touted by Black and Wiliam. The 2 x 2 experiment conducted by Martinez and Martinez involved 120 college algebra students randomly selected from 300 students placed in the remedial course as a result of their American College Test (ACT) mathematic scores. These 120 students were then randomly assigned to four classes of 30 students each. Two teachers, one experienced/ "excellent"(Teacher 1) and one less experienced/ "average" (Teacher 2) each taught a group of students using the "one-attempt testing" method and the "repeated testing" method. The control groups for the study were the classes that used the one-attempt testing technique. Using the same formula previously described, I calculated that the effect size for frequent testing compared to one-attempt testing for Teacher 1 was 0.43 and for Teacher 2 was 1.00. This indicated that there were no substantial effects for the type of teacher, but that the repeated testing yielded greater student achievement gains.

Although the findings of Martinez and Martinez emphasized the significance of frequent assessments on increased student achievement more so for novice teachers, only two teachers and 120 students were examined in the study. Absent from the discussion was the type of assessments used and details pertaining to the feedback provided once assessments had been administered. It is ironic then that Black and Wiliam (1998) would classify this as empirical evidence considering, according to them, these areas are two requirements of formative assessment (Dunn & Mulvenon, 2009a).

In Butler's (1988) experiment the focus was on formative evaluation feedback and the link to intrinsic motivation. The sample included 132 fifth and sixth grade Israeli students that were randomly assigned to one of the three experimental groups, therefore each experimental group was comprised of 44 students each. Using mathematics and language arts assessments, the students were selected randomly from both the top and bottom quartiles of their class to participate in the three session study. All students, both high achievers and low achievers were paired and assigned two tasks, one testing divergent thinking and the other testing convergent thinking. Students were then grouped by which type of written feedback they obtained, whether it was comments only, grades only, or grades and comments.

Black and Wiliam (1998) summarized that Butler's conclusions indicated:

> That even if feedback comments are operationally helpful for a student's work, their effect can be undermined by the negative motivational effects of the normative feedback, i.e. by giving grades. The results are consistent with literature which indicates that task-involving evaluation is more effective than ego-involving evaluation, to the extent that even the giving of praise can have a

> negative effect with low-achievers. They also support the view that pre-occupation with grade attainment can lower the quality of task performance, particularly on divergent tasks. (p. 5)

Although not directly reported in Butler's study, based on my calculations, effect size numbers appear to support Black and Wiliam's conclusions. I used "grades only" as the control when calculating effect sizes because it is the most common form of assessment. Focusing first on task A and sessions 1-3 respectively, when using the "grades only" group as the control and the "comments only" as the experimental group, the effect sizes for high achieving student results were -0.10, 0.04, and 0.83; whereas low achieving student results had effect sizes of 0.06, 0.68, and 0.77. When keeping the "grades only" group as the control and altering the experimental group to the "comments and grades" group, high achieving student results had the effect sizes of -0.01, -0.83, and -0.50; whereas low achieving student results had the effect sizes of -0.05, -0.76, and -0.43. These effect sizes determine that the formative evaluation feedback of "comments only" had a greater effect for low-achieving students, especially by the third session.

When calculating the effect sizes for the results of Task B for sessions 1-3 respectively, once again the "grades only" group acted as the control while the "comments only" as the first experimental group. Findings for the high achievers indicated effect sizes of -0.04, 1.43, and 1.51 and for low achievers -0.01, 0.85, and 1.88. When using the same control group, but "comments and grades" as the experimental group, high achievers effect sizes were -0.14, -0.26, and 0.01; low achievers effect sizes were -0.05, -0.11, and 0.22. For Task B, the divergent thinking test, once again the effect sizes determine that the formative evaluation feedback of "comments only" had a greater effect for low-achieving students, particularly by the third session.

The calculated effect sizes revealed that the "comments only" experimental group had the most significant gain in student achievement by the third session. This finding revealed that effect sizes were cumulative and by the third session students gained more by "comments only" feedback rather than by "grades only" and by "comments and grades". In fact, the results showed that when students were strictly focused on grade attainment, their performance on divergent tasks were negatively affected. The overall results of the calculations revealed that for both high and low achievers the ranking from most effective formative assessment to least was "comments only", then "grades only", followed by "grades and comments".

A concern acknowledged by the author as well as Black and Wiliam (1998) is that the tasks completed in the experimental groups were not components of or related to actual curriculum material. Couple that with the fact that graduate students, not the students' regular teachers, presented and conducted the experimental activities. The result is a study that lacks "ecological validity".

The experimental work of Schunk (1996) focused on two studies which investigated the affects goals and self-evaluation had on motivation and student achievement outcomes for fourth graders. Schunk hypothesized that learning goal orientation would result in increased student motivation and result in improved academic achievement more so than performance goals. Schunk explained that in both examined studies "students worked under conditions involving either a goal of learning how to solve problems (learning goal) or a goal of merely solving them (performance goal)" (p.359).

In Study 1, 44 9 and 10 year olds from one United States elementary school were randomly assigned to one of four different treatment groups: (a) learning goal with self-evaluation (LG-SE); (b) learning goal without self-evaluation (LG-NoSE); (c) performance goal with self-evaluation (PG-SE); and (d) performance goal without self-evaluation (PG-NoSE). Teachers for the instructional program were two female graduate students with what Schunk described as "some" teaching experience. Measures of goal orientation, skill, persistence, and self-efficacy were comprised and assessed on a pretest administered by an outside source to students. Over a seven school-day period, students worked on seven instructional packets; each packet representing one of the instructional sessions. The material included in six of the packets focused on each of the six major fraction skills, and the seventh packet contained review material. The four experimental groups were then assigned to two different treatments. In two of the four groups learning goals (how to solve problems) were stressed by the instructor; in the remaining two groups instructors emphasized performance goals (merely solving the problems).

Students participating in the LG-SE and PG-SE treatment groups evaluated their fraction capabilities in regards to problem-solving at the conclusion of each individual session, for a total of six evaluations. The evaluation material and procedures were duplicated from the pretest. The LG-NoSE and PG-NoSE treatment groups were instructed to complete an attitude questionnaire at the end of the first six sessions.

Whether conditions were learning goal or performance goal oriented, students that practiced self-evaluation performed better than those that did not use the specific formative assessment technique in posttest results measuring self-efficacy, skill, and persistence. Considering the necessary information was provided, I calculated the effect size for the LG-NoSE and PG-NoSE groups for the three aforementioned criteria. The effect size for the LG group using self-evaluation for self-efficacy was 0.26, for skill it was 0.21, and for persistence it was 0.68. The effect size for the PG group using self-evaluation for self-efficacy was 1.97, for skill it was 1.43, and for persistence it was 0.73.

Schunk's (1996) findings revealed that the PG-NoSE treatment group scored lowest on the outcome measures of self-efficacy, skill, and persistence than all other treatment groups. Therefore, according to Schunk and Black and Wiliams (1998), this was evidence indicating students who practiced the formative assessment technique of self-evaluation frequently, had greater motivation and increased achievement outcomes in comparison to those who did not participate in the practice of self-evaluation frequently.

The information obtained from Study 1 was further confirmed in Study 2. In Study 2 the sample was comprised of 40 9 through 11 year old fourth-graders. These 40 students were randomly assigned to either a learning goal (LG) or performance (PG) treatment group. Pretest and posttest procedures followed the same guidelines as Study 1, however in Study 2 all subjects participated in self-evaluation and were assessed only one time at the conclusion of the sessions as opposed to the six times in Study 1. The results indicated the LG treatment group scored higher than the PG treatment group on self-efficacy, skill, and task orientation.

Dunn and Mulvenon (2009a) acknowledged that Schunk's (1996) work was properly conducted. However, issues reside in the fact that both of Schunk's studies involved limited sample sizes, just 44 for Study 1 and 40 students for Study 2 (which later were divided into even smaller treatment groups). In addition, all subjects were selected from a total of four classes in just one elementary school. Black and Wiliam (1998) conceded that the work of Schunk was relatively sound, the material used for the study was curriculum based and applicable to a variety of teachers, however, they pointed out that the ecological validity of the study was closer in Schunk's work than the others previously discussed. The fact that self-evaluation took place as frequently as it did for experimental purposes is something that is unusual and not considered normal classroom practice. It is neither possible nor plausible to make grand assumptions on the positive impact of formative assessment usage based on such minimal data. This particular study demonstrated that it is important to identify specifically what type of formative assessment is valuable for increasing student achievement, rather than operating on assumptions that all types of formative assessments equally increase student performance.

The experimental study hailed by Black and Wiliam (1998) as one that "illustrates again the embedding of a rigorous formative assessment routine within an innovative programme" (p. 5) was conducted by Bergan, Sladeczek, Schwarz, and Smith (1991). The authors' work focused exclusively on investigating 838 disadvantaged kindergarteners in six different locations- Arizona, California, New Mexico, Iowa, Louisiana, and Mississippi. A total of seven school districts and seven rural and 14 urban schools were included in the sample.

The goal of the study was two-fold: (a) to examine the effects of a measurement and planning system (MAPS) and (b) to investigate the effects of MAPS on retention and special education placement referrals. The work of Bergan et al. (1991) was a component of a larger examination of the Head Start program, in which underprivileged students received free preschool funding. Their study tracked the Head Start students from preschool as they entered elementary school. I am inclined to acknowledge that the project was funded partially by the Head Start program and partially by a grant from the Ford Foundation.

Fifty-six teachers participated in the study (27 control and 29 experimental). This study did not focus predominately on one subject content area, but instead three areas of instruction; math, reading, and nature and science. Teachers placed in the experimental group received training on the implementation of MAPS. Two weeks into the program, after students had been working at their own individual level, students were assessed and adjustments were arranged after further diagnostic review and plans were modified based on individual needs after 4 weeks. The course lasted for an 8 week period.

The authors contended that their work had solidified the necessity of formative assessment implementation in order to assist with early math and reading basic skill acquisition. The authors reported students who attained basic skills early in their education development reduced the likelihood that they were placed in special education or remedial programs in the future. Bergan et al. (1991) reported that the results indicated that students whose teachers followed MAPS scored statistically significantly higher academically in developmental basic skills in math, reading, and science than those students of the control group that did not experience MAPS. However, the data collected from this study demonstrated a practically insignificant increase in student achievement results once calculations were completed in terms of effect sizes for the posttest mean differences (math 0.002, reading 0.04, and nature and science 0.12) although the authors stated otherwise. Even when adjusting for the scale score point disadvantages of the treatment group on the math pretest, the effect size is only 0.17 in favor of the treatment. Similar results follow for science at 0.15 and reading at 0.15. Due to the fact that formative assessments were a component of the newly implemented assessment program (MAPS), it is unclear as to how formative assessments individually would have affected student learning (Dunn & Mulvenon, 2009a).

In order to provide lower achieving and younger students the opportunity to enhance their understanding and knowledge of scientific inquiry, White and Frederiksen (1998) developed a middle school science-based, computer enhanced curriculum with the support of two experienced, certified educators. For the experiment, three teachers and 12 classes (a mix of seventh, eighth, and ninth grade) and an average of 30 students per class from two different schools participated. Instruction centered on teaching students about force and motion. Black and Wiliam (1998) recounted:

> All the work was carried out in peer groups. Each class was divided into two halves: a control group used some periods of time for a general discussion of the module, whilst an experimental group spent the same time on discussion, structured to promote reflective assessment, with both peer assessment of presentations to the class and self-assessments. (p. 6)

The control and experimental groups both included students with high and low Comprehensive Test of Basic Skills (CTBS) percentile scores. The analysis of project scores included a Mass Project (completed at the midway point) and the Final Project (completed at the conclusion of the program). The findings revealed that students in the experimental group using reflective assessment scored significantly better than students in the control group on the Mass Project, in fact the effect size for low-CTBS students was reported as 1.44.

It is evident that reflective assessment had a greater influence on students with low CTBS scores than on those with higher scores. For instance, the authors reported the effect sizes for low CTBS on the Mass Project assessment criteria (understanding, inquiry, connections, design, using tools, reasoning, communication, and teamwork) ranged from 0.25 to 1.03. High CTBS students did not fare as well; effect sizes ranged from -0.13 to 0.34 and were not as wide-spread amongst the criteria as the low-CTBS students. In regards to the Final Project, results further supported the effect of reflective assessment on student academic outcomes by an effect size of 1.70. Since CBTS scores "ranged from the 1st to the 99th percentile, indicating that the students' achievement levels approximate those of a national sample with a median percentile score of 60" (White & Frederiksen, 1998, p.29) this may have accounted for the variation in effect sizes between high and low CTBS students.

Meta-analysis.

Fuchs and Fuchs (1986) conducted a meta-analysis comprised of 21 "controlled studies" which yielded 96 various effect sizes investigating the effects of formative assessment on student achievement. The meta-analysis of Fuchs and Fuchs emphasized the effect of systematic formative evaluation, related primarily to special education students. Fuchs and Fuchs differentiated between systematic formative evaluation aptitude treatment interaction (ATI) by reporting:

> Whereas an ATI approach emphasizes the importance of describing salient learner characteristics, systematic formative evaluation focuses on ongoing evaluation

> and modification of proposed programs. Specifically, this approach employs regular monitoring of handicapped students' performance under different instructional procedures. The purpose of this monitoring is to provide a data base with which individualized programs may be developed empirically. Thus, systematic formative evaluation is an inductive, rather than deductive, approach to developing instructional programs. (p. 200)

Studies reviewed included both handicapped and non-handicapped students ranging from preschool to high school, more specifically 83% of the 3, 835 participates involved were mildly handicapped, accounting for 17% as non-handicapped.

An overall weighted effect size of 0.70 was reported for the 21 combined studies, calculated using Hedges's (1984, as cited in Fuchs & Fuchs, 1986) analogue to analysis of variance. An effect size of 0.70 "indicates that the upper 50% of the experimental group distribution exceeds approximately 76 % of the control group distribution" (Fuchs & Fuchs, 1986, p. 203). The effect size for non-handicapped students was reported as 0.63, still greatly significant (Cohen, 1977). However, researchers identified methodological problems with Fuchs and Fuchs's findings. Dunn and Mulvenon (2009a) expressed concern with the categorizing system Fuchs and Fuchs employed when labeling the quality of each potential study as good, fair, or poor. Dunn and Mulvenon recalled that out of the 96 reported effect sizes; eight studies were considered poor quality, 69 as fair, and 19 as good. It is necessary to review the criteria used to code each study included in the meta-analysis. Threats to validity were classified by raters as "serious" or "less serious" using the following guidelines:

> "Serious" threats included (a) un-equivalent subject groups, (b) confounded experimental treatments, and (c) nonrandom assignment of subjects to treatments. Examples of "less serious" threats were (a) the use of technically inadequate dependent measures, (b) uncontrolled examiner expectancy, (c) unchecked fidelity of treatment, (d) the employment of inappropriate statistical unit of analysis, and (e) inadequate teacher training. (p. 202)

A study comprised of at least one serious threat or less than four "less serious" flaws in design were labeled "poor quality". A study void of serious threats and a minimum of two "less serious" problems in methodologies were classified as "fair quality". Those studies representing no more than one "less serious" threat were considered "good quality". With that understanding, it is imperative to the validity of the meta-analysis to emphasize that 80% of the reviewed material included "research that was methodically unsound" (Dunn & Mulvenon, 2009a, p.5). In their review, Black and Wiliam (1998) questioned why with such significant evidence for the positive effect of formative assessment identified, would the work of Fuchs and Fuchs go almost unnoticed in future literature related to formative assessment practice.

Quasi-experimental studies.

Other articles referenced by Black and Wiliam (1998) that were used to draw conclusions on the positive impact of formative assessment were also jeopardized by the limited sample sizes. One such quasi-experimental study conducted by Whiting, Van Burgh, and Render (1995) elapsed over a span of 36 semesters, equivalent to 18 years. For their study Whiting et al. modeled the Bloom/Block formative assessment instructional strategy of mastery learning. The researchers credited "mastery techniques" and student self-awareness regarding personal learning styles for future academic success when engaged in "independent learning situations" (p. 12). Although 7, 179 students participated in the study and a wealth of information about formative assessment from end of the course teacher/course evaluations submitted by students and teacher feedback was obtained, all findings pertained to just one teacher's experiences. It is troublesome to decipher between the actual effects of the method of formative assessment practiced and the effects of the single studied teacher. The authors acknowledged that "some could argue that these results are produced by a gifted teacher who would be successful with any method. That may be true, however, we are convinced that mastery learning can make an excellent teacher outstanding, and certainly any teacher more effective" (p. 13). Nonetheless, Black and Wiliam (1998) reported that there were substantial gains in student achievement found in students that were enrolled in this particular teacher's class as opposed to those not taught by him, even though "the comparisons with the control group are not documented in detail" (p.4) and that it has been "reported that the teacher has had difficulty explaining his high success rate to colleagues" (p.4).

Black and Wiliam (1998) reported that the studies referenced in their work supported "conclusively that formative assessment does improve student learning" by producing student achievement gains that were "amongst the largest ever reported for educational interventions" (p.39). Be that as it may, many of the articles used to strengthen the argument in favor of the positive impact formative assessment usage has upon student achievement were laden with issues revolving around the research methodologies employed.

Web-based formative assessments—quasi-experimental studies.

In the 21st century not only has formative assessment use deluged the American classrooms, but its effectiveness as a digital data tool has also emerged. The development of recent web-based formative assessments have combated the predicament of one teacher attempting to "formatively assess" a classroom full of students in a timely and meaningful fashion (Wang, 2007). Wang's (2007) Formative Assessment Module of the Web-Based Assessment and Test Analysis System (FAM-WATA) was used to study 503 seventh-grade Taiwanese students at the direction of eight teachers with teaching experience in e-learning environments. A total of six effective formative strategies were analyzed: "repeat the test, correct answers are not given, query scores, ask questions, monitor answering history, and all pass and then reward" (Wang, 2007, p. 171).

Wang (2007) included a table labeled *Effectiveness of web-based formative assessment in e-learning* (p. 174) as evidence of the positive results associated with web-based formative assessments. The table lists eight studies and the findings of those studies. Phrases such as "helped to improve students' overall understanding" and "allowed individual students to monitor their educational progress" (p.174) were linked to/associated with the findings of the studies, however not a single effect size was reported nor was the information necessary to calculate effect sizes of the studies (means or standard deviations). Wang cannot justify his notion that evidence exists to claim web-based formative assessment positively influences e-learning environments if that information is omitted from his work. Therefore, it is hard to determine the validity of the claims purported by Wang.

Wang's (2007) findings concluded that use of FAM-WATA did in fact display great gains in student achievement levels in comparison to the group that used more traditional formative assessment strategies (i.e., pen and paper). Dunn and Mulvenon (2009a) criticized the implications Wang makes on the effectiveness of formative assessment. A control group of non-formative assessment users was not compared to groups that participated in formative assessment strategies, but rather compared with those that experienced different types of formative assessment, therefore generalizations regarding formative assessment benefits cannot be assumed.

Wang (2007) concluded his work proved "learning effectiveness will be enhanced if traditional paper and pencil tests are replaced by web-based formative assessment" (p.183) and moreover that using instructional programs and strategies similar to FAM-WATA will significantly enhance an already effective e-learning atmosphere. Although, he reported that pre and posttest scores were used in the data collection and analysis of his study, no information concerning posttest results were available in any presented tables, therefore I am unable to ascertain via calculations the effect size of the formative assessment interventions since none was reported. In addition, due to the lack of a true, non-formative, assessment control group, one is unable to decipher what the influence of formative assessment was when compared to a group that did not experience any type of formative assessment.

Sly (1999) conducted a study at the Curtin University of Technology located in Perth, Australia involving a year one economics course. Because the work of the study according to Sly was to contribute to student learning, it was therefore considered formative assessment. A total of 614 students made up the total sample for the study.

Results of students' scores on unit assessments that choose to take advantage of the practice test (n=417) were compared with the scores of students that did not utilize the practice test prior to the administration of the unit assessments (n=197).

For Sly's (1999) study a computer-managed learning (CML) system was used. Through CML the computer was the testing tool as well as the management tool, but never the teaching tool. Students that opted to take the pretest (experimental group) were to do so in the CML lab. This group participated in a practice test that assessed material that would appear in the first unit assessment. There was no available additional practice test for the second unit assessment. Students that did not take advantage of the CML system (control group) were only expected to take the unit assessments, forgoing the practice test.

Sly (1999) reported that information obtained via his study secured the notion that formative assessment assisted students in making academic achievement gains. Sly contended his findings revealed that subjects who chose to participate in the formative assessment practice test performed significantly better than those who opted not to participate.

A major contention of the study is the lack of randomization of the experimental and control groups. Like the Wang (2007) study, no information regarding the effect size or the necessary data to calculate it was reported in the Sly study. The self-selection process exercised in the study also threatened the validity of the study. Since the option to partake in the formative assessment CML program was at the discretion of the students, there is no way to differentiate between innate personalities and individual learning styles of all year one economic students.

Velan, Kumar, Dziegielewski, and Wakefield (2002) from the Department of Pathology at the University of New South Wales in Sydney, Australia discussed the effectiveness of using a software feedback formative assessment tool, Questionmark Perception, to promote student learning in their study. Velan et al. were advocates of on-line formative assessments because of the inconvenience that is caused by using the traditional "paper-based" approach. According to Velan et al. these limitations included: (a) the organization of students and assignments of a proctor or instructor to preside over the assessment; (b) it is too time consuming to individually meet with all students to provide personal feedback; and (c) the laborious task of ensuring question reliability and validity.

Undergraduate medical students used Questionmark Perception at "strategic timepoints" (Velan et al., 2002, p. 282) throughout the course. Immediately after students submitted their assessment, individual feedback was available for students to view. After receiving feedback students had the option of repeating each assessment multiple times. The first and last attempts of student's scores that opted to repeat assessments on multiple occasions were compared (n= 44). Velan et al. reported that students performed significantly better on their third attempt than the first and that this affirmed Questionmark Perception and similar web-based assessments were motivational tools and personalized student learning.

Nevertheless, like the previous two studies analyzed, there were methodological concerns present in the Velan et al. (2002) study. The problematic issues began with the limited sample size of just 44 students as well as the fact that, once again, vital information was absent, such as the pre and posttest scores. Analyzed student performances that were included in the study were not done so randomly like a true experiment, but instead were selected only if students opted to take the same assessment multiple times. Since the same exact assessment was administrated three times, it cannot be determined if the computer generated feedback was responsible for the supposed increase in student achievement or if multiple exposure over a short period of time played a role at some point as a result of repeated measures. It is therefore inappropriate that Velan et al. theorized "that all students made a genuine attempt to answer the questions in the assessment, and that they learnt from doing so" (2002, p. 282) if that evidence is not cogent.

The work of Buchanan (2000) was similar to that of Velan et al. (2002) in which the effectiveness of an on-line individualized formative assessment program was analyzed. Although Buchanan's work was more successful in attempting to prove the positive influence of web-based formative assessment by combining the work of two studies, there are also methodological issues that warrant acknowledgement.

Subjects in Study 1 were comprised from a cohort of 232 undergraduate psychology students enrolled in a level 1 course. In this study a web-based formative assessment program, Psychology Computer Assisted Learning (PsyCAL) was integrated as part of the course requirement. When using the PsyCAL package, it was mandatory for students to complete three set exercises, which included 11-15 multiple-choice questions per exercise. Subjects also had the option to utilize the two available supplementary revision exercises. The number of times students used PsyCAL and the results of the summative (end of course) assessment were the measured variables in this study.

From the 232 enrolled students, information was only obtainable and usable for 155 students regarding the use of PsyCAL. Out of the 155 identifiable users of PsyCAL there were 148 available exam grades to make the comparison between formative assessment usage and academic achievement. According to Buchanan, results identified a positive correlation between the number of PsyCAL uses and exam performance. However, the reported effect size is extremely small and insignificant at 0.03 (Cohen, 1977).

In Study 2 a cohort of 214 psychology students enrolled in a level 2 research and statistics module had the option to use the PsyCAL package during the course as a way to assess the "added value" it brought to the learning of students. Included in this PsyCAL package were five exercises containing 10 multiple-choice questions each. Buchanan (2000) conceded that students that took advantage of the PsyCAL package "performed significantly better than nonusers" (p. 198). However, only 16 identifiable users made up the sample size, which Buchanan acknowledged was small. Once again, no specific information was provided that would permit one to make calculations in order to obtain the effect size for this study.

Henly's (2003) study focused on the influence WebCT, a formative assessment program available for commercial purchase, had on student learning in a metabolism/nutrition unit in a dental science Australian program. According to Henly, this particular unit was one that students expressed difficulty with in previous years. The WebCT program provided the opportunity for students to determine if the inclusion of the web-based formative assessment program was useful to student learning and if it should be employed for future courses.

Students used the formative assessment program on a voluntary basis. Three different tests were created by faculty members, which included a variety of question types, such as short answer, true/false, multiple-choice, and matching. The number of questions ranged between 8-10 items. At the 5 week point and 2 weeks prior to the final course examination (week 20), the number of times students accessed the tests were monitored. Henly (2003) also compared the top and bottom 10% of students on the patterns of usage and overall performance on the three tests. As students completed items on the tests and submitted responses, if possible the questions were immediately scored and available for students to review. For questions that required a short answer response, the computer program provided a generic sample to use as an acceptable response. Although students had the option to repeat tests several times, the only scores recorded were that of the first assessment.

Enrolled in the program at the time of the study were 51 students. Summary pattern results revealed that students accessed Tests 1 and 2 more often than Test 3. It was also evident that by Test 3 students were no longer repeating the test like they had done for the first and second. Henly (2003) reported that by week 20 the majority of students had attempted Tests 1 and 2, however only half (26 of 51) had taken Test 3. Henly reported:

> Students in the top band accessed the formative assessment on an average more frequently than those in the bottom band (Table 3). In addition, all top-ranking students accessed the first two tests at least once, and all but one accessed Test 3 at least once (results not shown). In contrast, the low ranking group all accessed the first test, one student did not access Test 2 and only two students accessed

> Test 3. There was no significant difference in the marks achieved by the two groups of students in the first two tests. The number of students in the low ranking group who completed Test 3 was insignificant to allow a valid comparison of scores in that test. (p.120)

Dunn and Mulvenon (2009a) found flaws in Henly's study and suggested that the "study [c]ould have been improved by controlling for factors such as motivation, self-regulation, and poor performance" (p. 8). Similar to the Sly (1999) study, Henly's study reflected problems with the self-selection process used. Dunn and Mulvenon observed that both "the Sly(1999) and Henly (2003) studies have based their conclusion of the impact of formative assessments on the higher performing students, with limited evidence of their utility for these lower performing students" (p.8). Like other previously reviewed studies, the Henly study also lacked the necessary information to determine the effect size of the intervention. This study once again proved that it is not possible with a limited sample population, self-selection process, and inclusion of high performing students to make conclusive conclusions on the effectiveness of a web-based formative assessment approach for increasing student achievement.

Interim Assessments

As with formative assessment, there has been an increase in use of interim assessments by schools and districts in the United States in an effort to improve student achievement (Goertz, Olah, & Riggan, 2009). Dunn and Mulvenon (2009b) delineated the differences between formative and interim assessments:

> Formative assessment is defined as assessment used by teachers and students to adjust teaching and learning, as compared to interim assessment that informs

> policymakers or educators at the classroom, school, or district level about student achievement levels and curriculum effectiveness. Defining assessments in this fashion may create confusion for consumers of assessment products and literature. (p.5)

Perie, Marion, and Gong (2007) consider interim assessments to be:

> The assessments that fall between formative and summative assessments including the medium-scale, medium-cycle assessments currently in wide use. Interim assessments (1) evaluate students' knowledge and skills relative to a specific set of academic goals, typically within a limited time frame, and (2) are designed to inform decisions at both the classroom and beyond the classroom level, such as the school or district level. Thus, they may be given at the classroom level to provide information for the teacher, but unlike true formative assessments, they results of interim assessments can be meaningfully aggregated and reported at a broader level. As such, the timing of the administration is likely to be controlled by the school or district rather than by the teacher, which therefore makes these assessments less instructionally relevant than formative assessments. These assessments may serve a variety of purposes, including predicting a student's ability to succeed on a large-scale summative assessment...diagnosing gaps in a student's learning. Many of the assessments currently in use that are labeled "benchmark," "formative," "diagnostic," or "predictive" fall within our definition of interim assessments. (pp.1-2)

Goertz, Olah, and Riggan (2009) conducted an exploratory study in which they investigated the use of interim assessments, as well as the policies that support their use in the classroom. For the purpose of their study, Goertz, Olah, and Riggan (2009) also used Perie, Marion, and Gong's (2007) definition of interim assessments. Included in their study were 45 elementary school teachers (Grade 3 and 5) selected by a purposive sample of nine schools located in two Pennsylvania school districts. All data from their study involved interim assessments in mathematics for the 2006-2007 school year. The researchers purposely selected an urban and a suburban district to study in an effort to gather information on how "policy supports for assessment and instructional improvement function in these different environments" (p.2).

Guidelines set forth by the two participating school district leaders determined that interim assessments would be used as "teaching tools" and "expected teachers to use assessment results to reflect on their instruction, to discuss and share common problems and instructional solutions, and to provide remediation and enrichment during a dedicated period of time following the assessments" (Goertz, Olah, & Riggan, 2009, p.3). The researchers reported that their "study showed that interim assessments are useful but not sufficient to inform instructional improvements" (Goertz, Olah, & Riggan, 2009, p.8). In addition the authors uncovered minimal evidence indicating that the interim assessments they "studied help teachers develop a deeper understanding of student's mathematical learning—a precursor to instructional improvements. Most items in the assessments did not provide actionable information on students' misunderstandings" (Goertz, Olah, & Riggan, 2009, p.8).

Characteristics of interim assessments.

According to Perie, Marion, and Gong (2007) the best commercially interim assessment programs can:

1. Provide an item bank reportedly linked to state content standards.
2. Assess students on a flexible time schedule wherever a computer and perhaps internet connections are available.
3. Provide immediate or very rapid results.
4. Highlight content standards in which more items were answered incorrectly.
5. Link scores on these assessments to the scores on end-of-year assessments to predict results on end-of-year assessment. (p.14)

Perie, Marion, and Gong (2007) maintained that the purpose of an assessment determines whether it is classified as a formative assessment or an interim assessment. The authors emphasized that:

> If the purpose of these assessments is to enrich the curriculum, challenge the students to self-diagnose their own learning, provide insights into any misconceptions the students have, or provide additional professional development for the teachers, many of these types of assessment systems are woefully inadequate. (p.14)

What Perie, Marion, and Gong (2007) did find was that most commercially-produced interim assessment systems currently do not:

1. Address well multiple purposes, i.e., instructional, evaluative, or predictive.
2. Provide rich details about the curriculum assessed.
3. Help teachers understand the nature of a student's misconception(s).

4. Report detailed information on the student's depth of knowledge on a particular topic.
5. Further a student's understanding through the type of assessment task.
6. Give teachers the information on how to implement an instructional remedy. (p.14)

Perie, Marion, and Gong's (2007) contend that formative assessment activities differ from interim assessments in that they:

1. Are embedded within the learning activity and linked directly to the current unit of instruction.
2. Are small-scale (a few seconds, a few minutes, less than a class period) and short-cycle (they are often called "minute-by-minute" assessments or formative instruction).
3. Tasks presented may vary from one student to another depending on the teacher's judgment. (p.1)

If formative assessment and interim assessments are to be used in their true capacities and increase student achievement, and as a result reform and drive instructional practices at the grassroots level, a thorough understanding of the two evaluation systems is necessary.

Synthesis.

It is evident from the formative assessment studies included in this literature review that there are various and at times numerous, methodological concerns regarding the manner in which information is reported about the effectiveness of the treatment. From the eight "traditional" formative assessment studies discussed, when available or calculated, overall effect sizes ranged from insignificant to one in particular noteworthy meta-analysis with a gain of 0.70. For the web-based formative assessment programs, even less compelling information regarding effect size was exposed. Considering many policymakers and education leaders, in light of the NCLB Act regulations, view formative assessment as a means to a greater end, it may be surprising for some to learn of the insignificant conclusions discovered through this literature review.

Although the term formative assessment, also referred to as formative evaluation, dates back well before accountability policies existed, it is appearing more frequently in education policy documents as a result of recent accountability measures (Bell & Cowie, 2000), despite the lack of conclusive evidence supporting effects of formative assessment on student achievement. Stiggins (2002) addressed the "assessment crisis" prevalent in the American education system by posing the following thought-provoking questions that those who are responsible for improving student learning must consider:

> Are our current approaches to assessment improving student learning? Might other approaches to assessment have a greater impact? Can we design state and district assessment systems that have the effect of helping out students want to learn and feel able to learn? (p. 759)

A thorough review of the limited existing literature that attests formative assessment is influential in increasing student achievement, specifically maintained by Black and Wiliam (1998), has revealed severe to minor methodological issues that have now been exposed. It is critical that policymakers and education leaders do not fall prey to formative assessment practices based entirely upon the limited mixed reviewed highlighted empirical research, but rather focus on the questions raised by Stiggins.

By conducting a non-experimental study involving a modern, 21st century web-based formative assessment program with a sample size exceeding 600 middle school students, I can add valuable empirical information to the limited field of formative assessment and the implications it has on student achievement, which can assist policymakers, education leaders, and educational researchers with making informed decisions based on cogent evidence.

Formative assessment conclusions.

Only two of the eight "traditional" formative assessment referenced studies worked with middle school students, albeit none of the five reviewed web-based studies focused on middle school students. As a result, high-quality empirical evidence is essential, thus the contributions of my work are constructive on multiple levels.

With the year 2014 rapidly approaching and New Jersey state education funding drastically decreasing (Executive Order No.14, NJDOE, 2010a) it is vital that school district leaders use their allotted, minimal funds wisely. With academic stakes at an all time high due to the NCLB Act mandates, education leaders and teachers must provide said evidence of gains in student academic achievement. Accordingly, prior to purchasing a costly web-based formative assessment program without a surefire guarantee on an already strained budget, educator leaders and educators need a viable and evidence-based resource they can reference. The conclusions of my study aim to identify what the strongest predictors of student achievement are, and therefore the results could potentially serve as this invaluable resource.

Teacher Variables and Student Achievement

As far back as 1966, research indicated that "schools bring little influence to bear upon a child's achievement that is independent of his background and general social context" (Coleman et al., p. 325). Other researchers investigated the effects of class size (Finn & Achilles, 1990, 1999; Glass, Cahen, Smith, & Filby 1982; Jepsen & Rivkin, 2009; Mosteller, 1995;), teacher qualifications (Adams, Hutchinson, & Martray, 1980; Barnes, Salmon, & Wale, 1989; Darling-Hammond, 2000; Darling-Hammond, Hudson, & Kirby, 1989; Evertson, Hawley, & Zlotnik, 1985; Ferguson, 1991, 1998; Glassberg, 1980; Goebel, Romacher, & Sanchez, 1989; Gomez & Grobe, 1990; Jelmberg, 1996; Rivkin, Hanushek, & Kain, 2005; Strauss & Sawyer, 1986; Taylor & Dale, 1971), as well as other factors that might positively or negatively influence student learning. As previously addressed, there is debate surrounding the effect school resources have on student achievement. This section will further analyze teacher variables and their relationships to student mathematics and reading achievement.

The review of literature for teacher variables included in this section pertains to factors that can be measured (i.e., teacher education levels and certifications). I chose to include empirical studies published in peer-reviewed journals that pertained to these measurable teacher factors because the available literature relating to other intangible aspects of teachers (i.e., subject knowledge, intelligence, enthusiasm) are not empirically sound and suffer from methodological issues. For example, there are great variations as to what classifies as "effective" teacher characteristics. Most of the findings that have been reported on intangible teacher variables, such as measures of academic ability, measures of subject matter knowledge, and teaching knowledge, are greatly mixed

(Darling-Hammond, 2000). In addition, intangible factors are not reviewed because these aspects cannot be measured, collected, or placed into a regression analysis for my study. Rivkin, Hanushek, and Kain (2005) summarized my position most accurately:

> Prior investigations of school and teacher effects raised as many questions as they have answered, in large part because of the difficulties introduced by the endogeneity of school and classroom selection and in part because of the failure of observable teacher characteristics to explain much of the variation in student performance. (p. 449)

Research regarding National Board Certification for teachers was also omitted from this review. Due to the low number of New Jersey National Board Certified Teachers (NBCT) (currently 93 in the entire state) and none in the school district identified for this study, I chose to exclude literature related to NBCT and student achievement. Also, this will not be a factor to consider in the regression analysis, however with a high percentage of this DE District's teachers holders of alternate route certifications as well as master's degrees, I felt it necessary to include research relevant to these areas. Therefore, I focused solely on empirical research that related to teacher education levels and teacher certifications.

Using Darling-Hammond's (2000) review of previous literature pertaining to school inputs and student achievement, each teacher variable reviewed will include a section that summarizes the historical studies which will then be followed by more recent studies examining the same factors for predicting student achievement outcomes.

Teacher certifications and degrees.

Prospective teachers must complete a series of requirements prior to obtaining a standard teaching certificate. What those requirements actually consist of vary greatly from state to state. Generally, a standard certificate:

> Means that a teacher has been prepared in a state-approved teacher education program at the undergraduate or graduate level and has completed either a major or a minor in the field(s) to be taught plus anywhere from 18 to 40 education credits, depending in the state and the certificate area, including between 8 and 18 weeks of student teaching. (Darling-Hammond, 2000, p.7)

After completing the requirements issued by the student's college and the state, an examination (i.e., Praxis) must be passed in order to receive a standard teaching certificate. The majority, 85%, of new teachers entering the field of education today are graduates from traditional teacher preparation programs from across the United States (Boyd, Goldhaber, Lankford, & Wyckoff, 2007; USDOE, 2009).

Aside from obtaining a standard teaching certificate, states also award "emergency" and "alternate" certification routes toward licensure. These alternate routes allow those candidates that have not met all the requirements of teaching and that are unable to attain a standard license, the opportunity to get hired as a teacher. Goldhaber and Brewer (citing Shen, 1997) recalled "in just a 6 year period, from 1986-1992, the number of states allowing alternative certification jumped from 18 to 40" (2000, p. 131). At the time of this review, 47 states and Puerto Rico allowed alternative routes toward obtaining a teaching license. For the 2004-2005 academic year, 70% of alternate route program completers attended programs in just five states; Texas, New York, California, New Jersey, and Georgia (USDOE, 2009). It has been reported that more than a third of New Jersey, Texas, and California's new teachers are alternate route recruits (Boyd et al., 2007).

Overview of previous findings.

The question then becomes, do teachers that hold a traditional, standard certificate influence student performance more so than those teaching via the alternative route method? After reviewing literature on the topic in question, Evertson, Hawley, and Zlotnik (1985, as cited in Darling-Hammond, 2000) reported:

> The available research suggest that among students who become teachers, those enrolled in formal preservice preparation programs are more likely to be effective than those who do not have such training. Moreover, almost all well planned and executed efforts within teacher preparation programs to teach students specific knowledge or skills seem to succeed, at least in the short run. (p.8)

In her review of previous research conducted studying the effect of teacher preparation, Darling-Hammond (2000) recalled several studies that suggested "the typical problems of beginning teachers are lessened for those who have had adequate preparation prior to entry (Adams, Hutchinson, & Martray, 1980; Glassberg, 1980; Taylor & Dale, 1971)", and that "teachers admitted with less than full preparation—with no teacher preparation or through very short alternate routes—have found such recruits tend to be less satisfied with their training (Darling-Hammond, Hudson, & Kirby, 1987; Jelmberg, 1996)" (p.8).

Darling-Hammond (2000) recalled a study that Gomez and Grobe (1990) conducted in which Dallas-based alternative route candidates were examined upon entry to the classroom after completing a brief summer preparation training program. Once responsibilities were assumed, teachers were ranked on different aspects of teaching. Although observed teachers did receive average ratings on some facets of teaching, "they were rated lower on such factors as their knowledge of instructional techniques and instructional models" (Darling-Hammond, 2000, p. 8). Gomez and Grobe (as cited in Darling-Hammond, 2000) also reported that many more "poor" ratings were awarded to these alternate route candidates than traditionally trained teachers on the assessed teacher factors, from two to sixteen times as many in some cases. Perhaps the most compelling component of the Gomez and Grobe (as cited in Darling-Hammond, 2000) study was the reports that in language arts, student scores seemed to be greatly effected by the type of certification held by the teacher. Findings revealed that student's scores were significantly lower for those that had alternative route certified teachers than their peers who were taught by traditionally trained educators (Darling-Hammond, 2000). Citing Feiman-Nemser and Parker (1990), Gomez and Grobe (1990), Grossman (1989), and Mitchell (1987), Darling-Hammond, Berry, and Thoreson (2001) recalled that these studies proved that alternate route teachers "tend to have greater difficulties planning curriculum, teaching, managing the classroom, and diagnosing students' learning needs" (p. 69).

Goebel, Romacher, and Sanchez (1989, as cited in Darling-Hammond, 2000) found, after conducting an investigation of Houston's alternative route certification program that there was no association with the type of certification (traditional or alternative) that a teacher holds and student performance. However, Darling-Hammond (2000) called attention to the methodological issues raised regarding the researchers' study, identifying control concerns. Apparently, the study failed to control for initial test scores of the students reviewed and also lacked proper comparison groups when accounting for teacher experience. The work of Goebel et al. (1989) cannot soundly measure the effects of their study if the controls were not properly compared; "first year traditionally trained teachers were compared to two groups of alternative certification recruits, one with 1-4 years of experience and the other with 5-7 years of experience" (Darling-Hammond, 2000, p.9). Goldhaber and Brewer (2000) also cited Goebel et al. as a study that failed to prove a relationship exists between teacher certification route and student performance outcomes.

Another Texas study cited by Goldhaber and Brewer (2000) as one of the few that found "students of alternate-route teachers do at least as well as pupils of traditionally licensed teachers" (p. 132) was Barnes, Salmon, and Wale (1989). Barnes et al. reported results of two Texas school districts that found little disparities among traditional and alternative route certified teachers outcomes, however the report failed to provide any empirical data or methodology procedures adopted (Darling-Hammond, 2000). Darling-Hammond (2000) summarized the information provide by Barnes et al.:

> The study's table listing program types evaluated included 1 to 2-year university-based master's programs (which are called "alternative" in Texas because they are not undergraduate models) as well as district alternative programs that generally offer only a few weeks of summer training. In this case, the "alternative" group included programs providing extensive graduate level training along with those with very little preparation, thus preventing assessment of the effects of preparation on teacher effectiveness. (p. 9)

With insufficient data and a lack of controls, there is little basis for claiming that this study, along with Goebel et al.'s (1989) confirms or rejects an existing relationship between teacher certification routes and student achievement (Darling-Hammond, 2000).

Goldhaber and Brewer (2000) and Darling-Hammond (2000) contended that there are studies, Strauss and Sawyer (1986) and Ferguson's (1991, 1998) work, that provide evidence of a link between teacher licensing examination averages and student performance outcomes. North Carolina requires teachers to take a licensing test that measures subject matter and teacher knowledge known as the National Teacher Exams. Using statewide teachers' scores, Strauss and Sawyer (as cited in Darling-Hammond, 2000; Goldhaber & Brewer, 2000) found that increased student performance on state standardized assessments were related to average teacher performance.

In an effort to analyze school input variables and the effect they had on student performance outcomes, Ferguson (1991, as cited in Darling-Hammond, 2000) examined 900 Texas school districts. When controlling for student socioeconomic background and district characteristics, Ferguson concluded:

> That combined measures of teachers' expertise—scores on a licensing examination, master's degrees, and experience—accounted for more of the inter-district variation in students' reading and mathematics achievement (and achievement gains) in grades 1 through 11 than student socioeconomic status. (p. 9)

Ferguson (1998, as cited in Goldhaber & Brewer, 2000) found that increased student mathematic performance was a result of higher teacher's averages on the Texas licensing examination. Both Strauss and Sawyer (1986) and Ferguson (1998) arrived at similar conclusions regarding teacher performance and the effect that has on student performance.

Empirical evidence.

Merely having a driver's license does not make one a good driver, the same could be said for teaching, according to a study by Goldhaber and Brewer (2000). Goldhaber and Brewer aimed to add to what they considered little research available regarding the various types of teacher certifications and how differences in credentials may affect student achievement outcomes. The data source used by Goldhaber and Brewer was a national set of surveys known as the *National Educational Longitudinal Study of 1988* (NELS: 88) that included approximately 24,000 eighth grade students. In an effort to determine how teachers' credentials influenced student learning at the 12th grade level in mathematics and science, a series of surveys were administered from the 1992-1998 time period. Goldhaber and Brewer controlled for a variety of student background information (i.e., race/ethnicity, sex, family structure, and family income) as well as 10th grade test scores.

NELS: 88 is a database that linked detailed teacher information (i.e., race/ethnicity, sex, degree level, experience, certification, etc.) directly to the test scores of students that were enrolled in their classes. In their regression models Goldhaber and Brewer (2000) included what type of degree each teacher held, whether it be a master's degree, a higher education degree (Ph. D., MD, or D.D.S.) and/or an education specialist degree. Goldhaber and Brewer noted that they excluded the group of teachers with bachelor's degrees or less, being that 99% of public school teachers hold a minimum of a bachelor's degree (as cited National Center for Education Statistics, 1997). "Which type of math and science teaching certifications do you hold from the state where you teach?" was a question asked to 12th grade teachers on the NELS: 88 survey. Teachers could respond by selecting the following: "regular or standard," "probationary," "emergency," "private school certification," and "not certified" in subject (Goldhaber & Brewer, 2000). Goldhaber and Brewer explained:

> In out statistical models, we measure[d] the impact of certification type relative to those who hold standard certification in their subject. We cannot be certain of the extent to which definitions of certifications vary from state to state or how individual teachers interpret this question. (p. 133)

Recognizing that Goldhaber and Brewer (2000) were investigating the effects of teacher licensure on student achievement in public schools, their sample consisted of students that were enrolled in the 12th grade only; 3,786 students in mathematics and 2,524 students in science. There were 2,098 mathematics teachers included which 86% of them held a standard certificate, while 82% of the 1,371 sampled science teachers did. Ultimately, Goldhaber and Brewer asserted:

> We find that the type (standard, emergency, etc.) of certification a teacher holds is an important determinant of student outcomes. In mathematics, we find the students of teachers who are either not certified in their subject (in these data we cannot distinguish between no certification and certification out of subject area) or hold a private school certification do less well than students whose teachers hold a standard, probationary, or emergency certification in math. Roughly speaking, having a teacher with a standard certification in mathematics rather than in private school certification or certification out of a subject area results in a

> 1.3-point increase in the mathematics test. This is equivalent to about 10% of the standard deviation on the 12th-grade test, a little more than the impact of having a teacher with a BA and MA in mathematics. Though the effects are not as strong in magnitude or statistical significance, the pattern of results in science mimics that in mathematics. Teachers who hold private school certification or are not certified in their subject area have a negative (though not statistically significant) impact on science test scores. (p.139)

Darling-Hammond, Berry, and Thoreson (2001) believed Goldhaber and Brewer's (2000) study was riddled with methodological flaws, and therefore criticized their findings.

The Darling-Hammond et al. (2001) article was not an empirical study, but rather a critique on the methodology employed in the study in which Goldhaber and Brewer (2000) found that "teacher certification is pervasive, [and] there is little rigorous evidence that is systematically related to student achievement" (p. 141). Referring to Goldhaber and Brewer, Darling-Hammond et al. emphasized:

> The study's problematic conclusions derive not only from over-generalization based on tenuous evidence but also from a misunderstanding of how state certification systems operate; a failure to examine the available data on the emergency certified teachers in question (a large share of whom are similarly prepared to those with standard certification); and a neglect of much of the existing research in the field. The authors ignore methodological solid work that would lead to different conclusions about the effects of preparation, while referencing studies that are methodologically inadequate to support conclusions about the effects of preparation or certification. (p.58)

Darling-Hammond et al. also noted that Goldhaber and Brewer's study was funded by the Thomas B. Fordham Foundation, an organization that is in favor of the termination of teacher certification requirements.

In an attempt to debunk the notion that teacher certification plays no role in student achievement outcomes, Laczko-Kerr and Berliner (2002) conducted their own ex-post-facto archival research design to determine how student performances compare when students are taught by "under-certified" or certified Arizona teachers. To clarify what constitutes an Arizona "under-certified" teacher, Laczko-Kerr and Berliner explained that there were three possible classifications: (a) "emergency" (this group held bachelor degrees from accredited colleges, however they had little coursework completed in the field of education, and passed a criminal background check); (b) "temporary" (the researchers claim this classification is rarely used and was comparable to "emergency" certification; and (c) "provisional" (these candidates have had some degree of teacher education training, but fall short of the requirements necessary to obtain a standard certificate. For their study, Laczko-Kerr and Berliner compared all types of under-certified teachers with those that met all of Arizona's standard certification requirements criteria labeling those teachers as "certified".

The fully or "regularly" certified teachers included those that have met all of Arizona state's requirements, including: a bachelor's degree from an accredited institution, the completion of 45 hours of education coursework (elementary or secondary), received a passing score on the Arizona Educator Proficiency Assessment (AEPA), an understanding of the United States and state constitution, and clearance in a criminal background finger print analysis. Laczko-Kerr and Berliner (2002) noted that some of the under-certified teachers included in their study were participants of the "Teach for America" (TFA) alternative route program. Citing Darling-Hammond (1994), Laczko-Kerr and Berliner affirmed TFA aimed to "plac[e] energetic, bright, but unqualified teachers into poor, urban school districts" (p. 23).

The sample was comprised of 293, third through eighth grade teachers across five Arizona school districts that were hired either for the 1998-1999 school year or the 1999-2000 school year. There were a total of 159 certified teachers, while emergency, temporary, and provisionally certified teachers accounted for the 134 "under-certified" population. During the matching procedures, under-certified teachers and certified teachers were matched on the following criteria: certification status, grade level taught, and highest degree attained. Laczko-Kerr and Berliner (2002) detailed how they ensured pairs of both groups of teachers (certified and under-certified) were matched appropriately. Laczko-Kerr and Berliner recounted "matches were made using the following rules: 1) matches were first made within the school, 2) matches were made within the same school district, and 3) matches were made between similar school districts" (p.24). Since all districts in Arizona were required to administer the nationally norm referenced standardized Stanford Achievement Test, Ninth Edition (SAT 9), Laczko-Kerr also used these student results to compare teacher effectiveness.

When reporting the results, Laczko-Kerr and Berliner (2002) disclosed:

> A one-way analysis of variance (ANOVA) was conducted in which the independent variable was teachers' certification, while the dependent variable was the student achievement scores of these teachers as measured

> in Normal Curve Equivalents (NCE) for reading, mathematics and language in 1998-1999 and 1999-2000. (p.33)

Results confirmed that students taught by certified teachers in 1998-1999, outperformed students taught by under-certified teachers. The differences were found to be statistically significantly higher for the reading and language tests. The authors indicated that although the results for the mathematics test were not statistically significantly higher among students taught by certified or under-certified teachers, the results did emulate those of the reading and language tests. For the 1999-2000 time period, students taught by certifies teachers outperformed students taught by under-certified teachers on all tests, reading, language, and mathematics at significant measures.

Laczko-Kerr and Berliner (2002) explained that the Normal Curve Equivalents (NCE) allow for certified and under-certified teacher evaluation differences. It was determined in reading that certified teachers outscored under-certified teachers by 6 NCE points, by 3 NCE points in mathematics, and approximately 5 NCE points in language for the 1998-1999 time period. For the 1999-2000 time period, differences in NCE points still favored certified teachers by 3 points in reading, 5 points in mathematics, and 2 points in language. In terms of effect size "these differences range across two years from .14 to .28 in reading, .14 to .24 in mathematics, and .09 to .19 in language" (Laczko-Kerr & Berliner, 2002, p.36). Due to departmentalization concerns in grades 7 and 8, that data was analyzed separately, however greater effect sizes were reported over the same two year span; "from .19 to .38 in reading, .24 to .28 in math, and .14 to .33 in language" (Laczko-Kerr & Berliner, 2002, p.36). Laczko-Kerr and Berliner (2002) concluded:

> That the average ES across all sub-tests of the SAT 9, across both years of testing, and across analyses, is around .20. Because of the relationship between effect size (ES) and yearly progress on standardized (Glass, 2002), one could expect that during one academic year in the primary grades, the students of certified teachers would make approximately 2 months more academic growth than would the students of under-certified teachers. (p.36)

Laczko-Kerr and Berliner (2002) raised awareness to how the 20% less academic growth due to under-certified personnel in the classroom is detrimental to at risk children whom were already low achieving prior to being placed in an already comprised learning environment. When considering the TFA alternative route program, Laczko-Kerr and Berliner (2002) found no significant differences among their students' scores in comparison with other under-certified teacher's students' performance and concluded "the TFA teachers are no better able to teach than any other under-prepared teacher" (p.41).

Teacher years of experience.

Teacher experience is another possible predictor of student achievement that stems from teacher certification issues (Laczko-Kerr & Berliner, 2002). For decades researchers have attempted to analyze the relationship between teachers' years of experience and student achievement. Some researchers discovered a positive link between a greater number of years teaching and gains in student achievement (Fetler, 1999; Hanushek, 1972; Hawkins, Stancavage, & Dorsey, 1998; Klitgaard & Hall, 1974; Murnane, 1975; Murnane & Phillips,1981; Rivkin, Hanushek, & Kain, 2005; Rowan, Correntti, & Miller, 2002), albeit at statistically insignificant levels. While others have determined that there is no significant relationship between teacher experience and student learning (Hanushek, 1971; Link & Ratledge, 1979).

Laczko-Kerr and Berliner (2002) highlighted some of the evidence that pointed to a relationship between teacher experience and student achievement reported by Hawkins, Stancavage, and Dorsey (1998). Hawkins et al., using 1996 NAEP analysis data reported "students who were taught by teachers with less than 5 years of teaching experience performed below the level of those students whose teachers had 6-10 years or 25 or more years of experience" (Laczko-Kerr & Berliner, 2002, p.13). Also discussed by Laczko-Kerr and Berliner were the findings of Lopez (1995). Using a large Texas data set, Lopez reported that in order for teaches to maximize their students' performance on tests, at least seven years of teaching experience is required.

Empirical evidence.

For their empirical analyses, Rivkin, Hanushek, and Kain (2005) used the data complied from UTD Texas Schools Project (which John Kain, one of the researchers, created and directed). The researchers claimed this rich data set could assist in identifying what school and teacher effects contributed to student achievement. Data throughout the mid-1990s from three cohorts was used for their study. Data from one cohort consisted of student test scores from grades 3 through 7, while the other two cohorts used grades 4 through 7 data. Rivkin et al. explained that each cohort was comprised of more than 200,000 students, drawn from over 3,000 public elementary and middle schools. The authors alleged their extensive sample size "permit[ted] much more precise estimates of school average test scores and test score gains" (p. 431).

Using student test scores from the Texas Assessment of Academic Skills (TAAS), the researchers focused primarily on mathematics and reading results. The researchers asserted that due to different reporting systems, they were unable to directly link student scores with specific teachers. Instead, the researchers used the teacher's personnel data (i.e., experience and highest degree earned) as well as school factors (i.e., class size, subject, and grade) "to construct subject and grade average characteristics for teachers" (p. 432) to link with student results.

Rivkin et al. (2005) professed throughout their report that "consistent with prior findings, there is no evidence that a master's degree raises teacher effectiveness. In addition, experience is not significantly related to achievement following the initial years in the profession" (p. 419). Rivkin et al. also declared "there has been no consensus on the importance of specific teacher factors, leading to the common conclusion that the existing empirical evidence does not find a strong role for teachers in the determination of academic achievement and future academic and labor market success" (p. 419). However, Rivkin et al. did not cite any previous evidence (empirical or not) to support these strong accusations.

In their conclusions, Rivkin et al. (2005) summarized:

> 1. Similar to most past research, we find absolutely no evidence that having a master's degree improves teacher skills.
>
> 2. There appear to be important gains in teaching quality in the first year of experience and smaller gains over the next few career years. However, there is little evidence that improvements continue after the first three years. (p. 449)

Rivkin et al. acknowledged that although it is natural to assume that in order to improve quality, teacher standards must be raised, however this is not a practical measure policymakers should take. The authors contended that they have added evidence to the notion that a teacher's education level and certification status does not equal quality. and that state officials should focus on enforcing "effective hiring, firing, mentoring. and promotion practices" (p.450).

Michel (2008) conducted a study using NJ ASK4 mathematics and language arts scores to determine what variables (student, school, and teacher) were the strongest predictors of student performance. Using a vast sample of 888 New Jersey public schools, including 72,267 grade 4 tested students and their mathematics and language arts scores, as well as various student (mobility rate, attendance rate, suspension rate, and expulsion rate), school (DFG, class size, length of school day, and faculty attendance rate), and teacher (percentage with National Board of Standards certificate, percentage with a master's degree. percentage with doctorate degree. and faculty attendance rate) variables published on the NJDOE website, Michel ran multiple regression analyses. The results of Michel's study identified that at all levels, partially proficient, proficient. and advanced proficient in both mathematics and language arts achievement, the strongest predictor of student performance was socio-economic status (measured by DFG).

When controlling for student and school variables, Michel reported that a significant predictor of student performance at the partially proficient and advanced proficient level in math and at all levels in language arts was the percentage of teachers holding a master's degree. Michel reported a positive relationship between student performance on the NJ ASK4 and increases in school percentages of teachers with a master`s degree, however he did mention that the relationship was rather weak.

Synthesis.

Although the research is mixed, much does indicate the importance that teacher education matters in specific subject area and education coursework and the effect that has on student achievement (Denton & Lacina, 1984; Ferguson & Womack, 1993; Guyton & Farokhi, 1987; Michel, 2008; Monk, 1994). Students need teachers in the classrooms, how they earn their certification can vary. What we cannot afford to do is jeopardize the future of our youth by permitting those that are not competent in the areas that they are responsible for and expected to teach our children. Research demonstrates that when teachers have taken exams that measure their understanding of specific subject content knowledge coupled with knowledge of teaching and pedagogy, effects to student achievement have been greater (Ferguson, 1991; Fetler, 1999; Hawk, Coble, & Swanson, 1985; Strauss & Sawyer, 1986).

It is hard to contest that schools strive to educate our students. Exactly how to achieve that and with what effective resources is the ongoing debate. It is imperative that policymakers review the evidence in the extent literature that reveals flaws in some of the ways in which teaching procedures are currently addressed. Through modifications to the manner in which teachers are awarded teaching licenses, to the preparation that is provided on site once employment is secured, policymakers must refocus the attention to the bodies of students in the classrooms, and not just getting "anybody" that has a certificate to teach our children.

Theoretical Framework

Introduction

There is an ongoing debate concerning which variables influence student achievement most significantly. Previous research has identified several types of "input" variables that influence student academic achievement. They can be categorized as school, student, and teacher variables. For the purpose of my study, the "output" variables were students' achievement on the NJ ASK 8 language arts and mathematics sections.

Variables that influence student achievement are generally categorized as either pertaining to the school, the student, or the teacher. Which variable influences have a statistically significant influence on student achievement than others? The answer depends upon the particular research results one consults. Some researchers reported that schools have very little influence on student achievement when socioeconomic status is held constant (Averch et al., 1974; Coleman et al., 1966; Jencks et al., 1972) whereas others disagreed; citing that teacher qualification greatly influences student academic achievement (Darling-Hammond, 2000; Ferguson, 1998).

What is not debatable is the influence, for better or for worse that NCLB mandates have had on student achievement. Regarding accountability protocol, Paulson and Marchant (2009) recounted how standardized testing "has been heralded as *the* universal tool" (p.3) for measuring this. In order for the standardized test results to suffice as the major measure of accountability Paulson and Marchant (2009) emphasized the following assumptions must be:

> (a) that the tests reflect important standards of learning that are being taught in the schools; (b)that student who do not reach proficiency are inadequate in their knowledge and skills, regardless of their performance on other forms of assessment; (c) that these tests are better indicators of students' ability than the judgment of teachers; (d) that the collective scores of teachers' students reflect the quality of their instruction and it assumes that the collective scores of schools and districts reflect the quality of their educational programs; and (e) that the collective scores of test-takers from a state represent the quality of education and educational policies of the state. (p. 3)

Although the statewide tests administered in New Jersey schools might not meet all of the above criteria, the NJDOE has none-the-less determined that the NJ ASK standardized statewide test is the primary measure used for accountability purposes. The NJDOE (2010b) personnel, through the use of the New Jersey School Report and various other mandates developed a set of input variables that they claim influence student achievement. In essence they created a theoretical framework that supports their use and mandate of specific input variables as a method to raise achievement on their primary output variable, the NJ ASK. As stated in Chapter I, this study will explain the influence of input variables, identified by the state and found in the empirical literature to have a statistically significant influence on student achievement in past studies.

No Child Left Behind Act

The *NCLB Desktop Reference* published in 2002 identified the accountability requirements for schools according to the USDOE. The report explained:

> The NCLB Act is designed to help all students meet high academic standards by requiring that states create annual assessments that measure what children know and can do in reading and math in grades 3 through 8. These tests, based on challenging state standards, will allow parents, educators, administrators, policymakers, and the general public to track the performance of every school in the nation. Data will be disaggregated for students by poverty levels, race, ethnicities, disabilities, and limited English proficiencies to ensure that no child—regardless of his or her background—is left behind. The federal government will provide assistance to help states design and administer these tests. (pp.9-10)

Therefore, a greater focus on assessment results, specifically, statewide standardized test results can be attributed to NCLB mandates. In an effort to ensure that all children "reach proficiency on challenging state academic standards and assessments" (NCLB Desktop Reference, 2002, p. 13) the NJDOE administers the NJ ASK in grades 3-8 and the HSPA in grade 11. Some states, like New Jersey, took the idea one step further and recommended that the use of formative assessment was a school level variable that influenced student achievement. During the 2008-2009 school year, 1 year after introduction of FAT, the NJDOE endorsed and recommended computerized formative assessment tool, Commissioner of Education Lucille Davy detailed the NJDOE's stance on formative assessment in an memo addressed to all district administrators. Davy (2008) explained that:

> Formative assessment resources allow educators to evaluate and measure student achievement continually, in a low-pressure context, using non-secure benchmark testing forms, item pools, distractor analysis, item authoring software, and

> associated score reports. Formative assessment resources allow teachers to connect specific grade level indicators with specific students or groups of students. (p.1)

Davy continued discussing the need for formative assessment by promoting the corporate FAT product. Davy wrote:

> Teachers naturally want to see the test questions their students got wrong or right, and the [FAT- pseudonym] resources make that possible, well before those students sit down for the high stakes testing in the spring. Furthermore, the supporting professional development programs constitute an intellectually rich foundation for teachers who want to integrate a wide range of assessment practices and concepts into their regular instructional routines. We believe that using these formative assessment tools will in itself constitute professional development for teachers, but the formal workshops and web-based supports will assure that teachers are confident about the underlying pedagogy of formative assessment, not just the technology. (pp.1-2)

Whether the NJDOE personnel knew it or not, the moment they released this memo, they created another NJDOE recommended input variable. Since then, the NJDOE personnel have mandated the use of FAT in districts that have been awarded certain types of NJDOE competitive grants. Most recently all the districts awarded the four-year INCLUDE grant must also use FAT with their students.

The literature regarding the influence schools have on student achievement varies tremendously because it is difficult to measure what "school resources" encompasses. Previously discussed was the influx of formative assessment tools in classrooms. To add to the little empirical evidence on how these particular tools influence student achievement at the middle school level, my study determined the strength of the relationship between the "input" of the school formative assessment resource and the "output" of the performance of students on the NJ ASK6, 7, and 8. The formative assessment tool touted by the NJDOE is The Company product FAT.

Some have argued that throwing money at schools in an effort to increase student achievement is foolish (e.g., Hanushek, 1971, 1972, 1979, 1981, 1986, 1989, 1991, 1994, 1996, 1997), while others believe "that simply throwing money at the schools is an effective strategy for improving education" (Baker,1991, p.630). Although in the case of my study, money was not technically distributed by the NJDOE, it was encouraged that districts implement the web-based formative assessment tool FAT. In an effort to do so, FAT was provided free of charge for 5 years to NJ school districts with the hopes that its value would be recognized and after the trial period subsequently purchased. To date, FAT is in Year 4 of the 5 year trial period and had a total of 250 NJ school districts slated to implement the computer-based formative assessment system this year. However, there were severe budgets that only permitted 178 NJ school districts to implement FAT for Year 4. FAT representatives confirmed that during Year 1 of the trial period 50 NJ school districts had employed the formative assessment program. During Year 2 of the trial period 175 NJ school districts had implemented FAT, and 205 in Year 3.

Although the NJDOE encourages school districts to use the web-based FAT product to identify areas of weakness in students' learning, research reveals little is known at the middle school level how this input variable enhance or hinders student academic achievement. Research does reveal that particular types of traditional formative assessment (e.g., self-evaluation and systematic formative evaluations) have a slightly stronger influence on student achievement than other forms.

Teacher qualification is another input variable identified by the NJDOE personnel and USDOE personnel as having an influence on student achievement. According to the USDOE (2002), requiring that all core academic teachers (i.e. English, reading or language arts, mathematics, science, foreign languages arts, history, and government, economics, arts, history, and geography) are "highly qualified" is one way "to help ensure that all children have the opportunity to obtain a high-quality education and reach proficiency on challenging state academic standards and assessments" (p.13). The USDOE recognizes that "Highly Qualified Teachers" (HQT) as those that "have state certification (which may be alternative state certification), hold a bachelor's degree, and have demonstrated subject area competency" (p.19).

Nevertheless, the extant literature reviewed on the influence of teacher qualifications and credentials revealed mixed results. Although the USDOE and the NJDOE require core academic teachers to be "highly qualified," this is merely a label. HQT does not translate into "good quality" teaching. Research does not fully support that the "input" of HQT leads to the "output" of increased student achievement on state mandated tests, regardless of sanctions placed by the state and federal agencies. However, research has demonstrated that when teachers have taken exams that measure their understanding of specific subject content knowledge coupled with knowledge of teaching and pedagogy, effects to student achievement have been greater (Ferguson, 1991; Fetler, 1999; Hawk, Coble, & Swanson, 1985; Strauss & Sawyer, 1986). Regardless of the mixed research results, the NJDOE and USDOE personnel include

HQT in their theoretical framework as an input variable that influences student achievement and all districts in New Jersey must comply and make this input variable part of their plans to raise student achievement

In 1975 the NJDOE recognized that not every community in NJ had the ability to support public education at the same monetary levels. As a result, the District Factor Grouping system (DFG) was introduced to monitor equitable spending provisions and to provide the opportunity to conduct fair analyses of district-to-district test score results. As schools are labeled according to their DFG, the potential inequities brought on by various degrees of poverty are uncovered (Tienken, 2008a). Through NCLB (2002) regulations "Title 1 provides flexible funding that may be used to provide additional instructional staff, professional development, extended-time programs, and other strategies for raising student achievement in high-poverty schools" (p.13). Although the USDOE, via Title 1 funds has allocated billions and billions of dollars since its inception to improving student achievement, it is recognized that "the academic achievement gap in this country between rich and poor, white and minority students, remains wide" (p.9). While it is recognized as *an* influence by the USDOE and NJDOE, literature revealed that SES and family background remains *the* strongest predictor of student achievement (Averch, Carroll, Donaldson, Kiesling, & Pincus, 1974; Coleman et al., 1966; Gamoran & Long, 2006; Jencks et al., 1972; Michel, 2008; Smith, 1972; Tienken, 2008a).

It is important to note the dangers associated with relying solely on the DFG determined by the State. It is important to look at SES at the school and student levels using free and reduced lunch status. Case in point, the district used in this study is classified as a DE district; however the percentage of economically disadvantaged students ranges from 13.5% to 44.5% in one grade level throughout the district's five middle schools. This particular district that falls under the auspices of the DE DFG in actuality is a conglomeration of several DFGs ranging from the wealthy to the severely disadvantaged. New Jersey school districts are classified at the district level rather than the school level. This classification system becomes flawed when large discrepancies among student population concerning SES are evident as in the case of this study. Although the district services several hundred thousand students, it is not appropriate to simply apply a one-size "label" and assume all interventions will yield similar results in student achievement across the district.

Taken together, the USDOE and NJDOE personnel have identified, and in some cases mandated, specific input variables that they stated create as a set of variables that influence student achievement. A review of the empirical literature found support, albeit mixed in some cases, for three specific variables. Although not mentioned specifically in the empirical literature, the FAT product is a form of formative assessment, it is mandated in some districts in NJ by the NJDOE, and it is mandated for use in the researcher's district. Therefore, it was included in the initial conception of a theoretical framework to guide this study.

Further analysis of district specific characteristics of the initial variables included in this framework resulted in the elimination of teacher advanced degree status. This variable was later eliminated, because, at the school level, there was little to no variance in the degree status of the language arts and mathematic teachers. Thus, the variable was deemed moot. An additional possible influential variable was identified, Academic Support Instruction (ASI), due to the nature of its use in the DE District. In an effort to "specialize instruction based on the student's individual needs" (Woodbridge Township. 2008-2009) the DE District provides ASI in both language arts and mathematics in Grades 6-8. The DE District's Middle School Program of Studies Guide reported for ASI Mathematics that:

> Students will be required to enroll in this course if they scored below the designated levels of proficiency in the state's NJ ASK – 5, 6, or 7. Input from classroom teachers, guidance personnel, and building principals is also considered before final placement in this program is recommended. The content of this course includes topics from the appropriate grade-level mathematics curriculum as well as specialized instruction based on the student's individual needs. (p.11)

In ASI Language Arts it was reported that:

> Students will be required to enroll in this course if they scored below the designated levels of proficiency in the state's NJ ASK – 5, 6, or 7. Input from classroom teachers, guidance personnel, and building principals is also considered before final placement in this program is recommended. Courses Include: ASI Reading/Writing Workshop (Grade 6) and ASI Language Arts Workshop (Grades 7 & 8). The content of these courses includes topics from the regular grade-appropriate Reading/Writing and Language Arts courses as well as specialized instruction based on the student's individual needs. (p.11)

ASI Mathematics and ASI Language Arts courses are conducted during the regular school day and replace student's mandatory mathematics and language arts classes. ASI is not an after school or pull-out program. The inclusion of the ASI eligibility variable can provide useful information in regards to predictors that influence student achievement. School administrators exert direct control over entrance criteria, teachers assigned to teach the program, and curriculum. It is a variable that, if statistically significantly related to student achievement in some way, school administrators can mutate in an attempt to improve student achievement in the future. It is valuable for community stakeholders to ascertain if individualized instruction courses similar to the DE District's ASI program truly improve student achievement.

Figure 2 presents the final conceptual framework used to guide this study. Figure 1 depicts the theoretical framework used to guide the study.

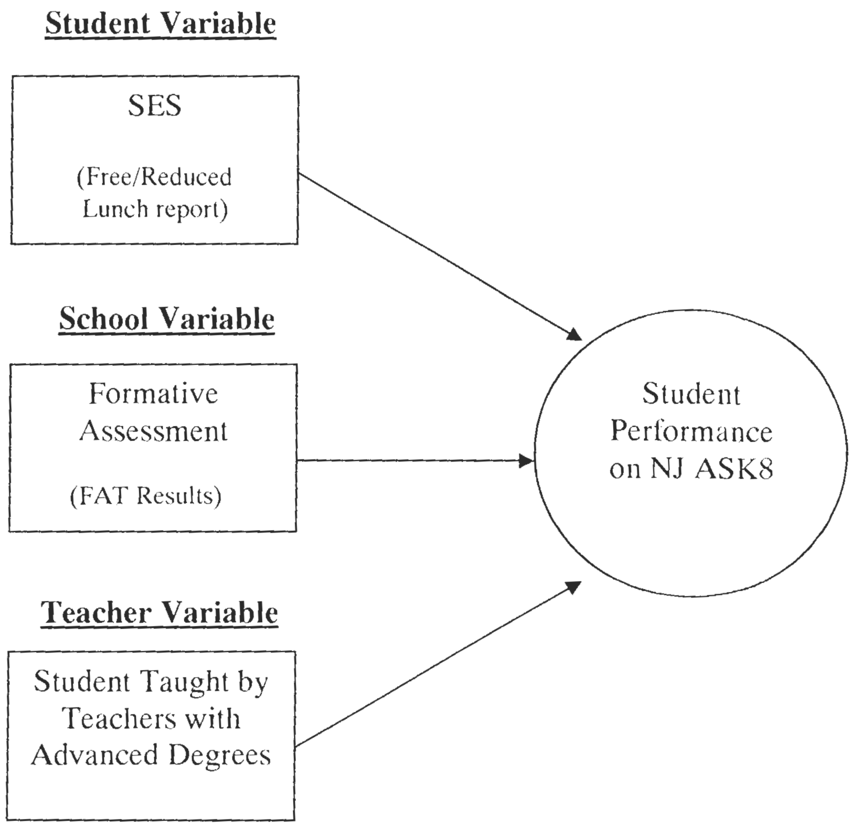

Student Variable
SES
(Free/Reduced Lunch report)
School Variable
Formative Assessment
(FAT Results)
Teacher Variable
Student Taught by Teachers with Advanced Degrees
Student Performance on NJ ASK8

Figure 1. Input/output framework.

Production Function Theory

The NJDOE uses a production function theory-base to make the above stated recommendations regarding the school and teacher variables that influence student achievement. When describing production function in the realm of higher education Hopkins (as cited in Hoenack & Collins, 1990) explained it is "intended to represent the process by means of which an institution—here, a college or university—transforms inputs (typically labor and capital) into outputs" (p. 11). For the purpose of this study, the institution becomes the school or school district, the inputs become the student, school, and teacher variables previously addressed and the output becomes the students' NJ ASK scores, which according to the USDOE and NJDOE represent student achievement levels.

Conclusion

Although the USDOE and the NJDOE personnel recognized that there are particular factors that influence student achievement, there are sometimes disconnects between the consistency in which these variables are found in the empirical literature to influence achievement and what NJDOE and USDOE personnel espouse. This study will add empirical evidence to the mixed, limited literature regarding the influences on student achievement evident in a society where accountability is at the forefront of many reform policies. This study could greatly benefit school administrators, educators, curriculum leaders, parents, school boards as well as education researchers in determining what impact, if any formative assessment has on student achievement. The study into these uncharted waters will also add to the current limited empirical evidence available in the literature to either support or refute the positive implications associated with formative assessment. The work of this study will showcase the detrimental influence to policymakers and district leaders of relying solely upon state test results to take action. The value and information gained from state test results is only as significant as the observer desires it to be. It is crucial that district leaders and administrators acknowledge the irreversible potential harm of exclusively using state test results for high-stakes purpose and the misfortune this may inflict upon particular students of the community.

Chapter III

METHODOLOGY

Introduction

The purpose of the quantitative study was to investigate the influence of student and school variables found in the extant literature on student achievement and aggregate district student NJ ASK scores in Grade 8 language arts and mathematics. By placing the focus on multiple student and school variables that have a statistically significant relationship to student achievement, this study aimed to produce research-based evidence to assist all stakeholders in public education regarding reform initiatives. A dearth of empirical evidence exists regarding student and school variables and the relationship of these variables to middle school student achievement on the NJ ASK8 language arts and mathematics sections. Therefore this study could add empirical results to the limited body of existing literature.

Research Design

"Non-experimental research is frequently an important and appropriate mode of research in education" (Johnson, 2001, p.3) due largely in part to the inability to perform randomized experiments and quasi-experiments. I conducted a non-experimental, cross-sectional, explanatory study. The correlational study only collected data from one point in time. Under the auspices of Johnson (2001) an explanatory study must meet the following criteria: (a) Were the researchers trying to develop or test a theory about a phenomenon to explain "how" and "why" it operates? (b) Were the researchers trying to explain how the phenomenon operates by identifying the causal factors that produce change in it? (p.9).

In order to determine which student and school variables had a statistically significant relationship to student achievement, I used simultaneous multiple regression models for my study. This strategy is used when the researcher has no logical or theoretical structure of the data. This method is typically used to explore and maximize prediction (Pedhazur, 1997). Scatter diagrams of residuals and normal probability plots of residuals were conducted to test assumptions.

Chapter II presented literature that indicated specific variables that are related to student achievement. However, to what extent these variables related to student performance on the NJ ASK8 at the middle school level is unknown. Because I did not know which variables would create the best prediction equation simultaneous regression was an appropriate method to use. Researchers use simultaneous regression when they have a limited number of predictors and are unsure of which variables would create the best prediction equation model (Leech, Morgan, & Barrett, 2008).

The results of this study came from school level data obtained from one school district with a District Factor Group (DFG) of DE. Generalizations cannot be made that similar results would prevail in districts located in wealthier communities. This study focused on one commercially produced standardized formative assessment product and only variables identified previously in the literature that influence student achievement on commercially prepared standardized tests: (a) eligibility for free and reduced lunch; (b) teacher degree status; and (c) formative assessment usage. The school variable of participating in the Academic Support Instruction (ASI) program, which is funded by Title I monies at the elementary level but not at the secondary level in the DE District, was also used. ASI will be included as a variable because the DE District has identified it as a factor that influences student achievement. Students are compelled to participate in this remedial instruction program based on the previous year's NJ ASK language arts and mathematic scores.

Research Questions

This study was guided by the following overarching research question: What student and school variables, found in the extant literature explain the greatest variance in student achievement on the NJ ASK8 language arts and mathematics sections?

Research Question 1: What is the strength and direction of the relationship between the proficiency categorizations of students in Grade 8 measured by commercially produced standardized formative assessment in language arts and the proficiency categorizations of students in Grade 8 measured by the language arts portion of the NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Research Question 2: What is the strength and direction of the relationship between the proficiency categorizations of students in Grade 8 measured by commercially produced standardized formative assessment in mathematics and the proficiency categorizations of students in Grade 8 measured by the mathematics portion of the NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Research Question 3: What are the statistically significant student variables that explain the largest amount of variance in student language arts achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Research Question 4: What are the statistically significant student variables that explain the largest amount of variance in student mathematics achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Research Question 5: What are the statistically significant school variables that explain the largest amount of variance in student language arts achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Research Question 6: What are the statistically significant school variables that explain the largest amount of variance in student mathematics achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Null Hypotheses

Null Hypothesis 1: No statistically significant relationship exists between a commercially produced standardized formative assessment tool proficiency categorization and students' language arts proficiency categorization on the NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Null Hypothesis 2: No statistically significant relationship exists between a commercially produced standardized formative assessment tool proficiency categorization and students' mathematics proficiency categorization on the NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Null Hypothesis 3: There are no statistically significant, research demonstrated, student variables that predict student language arts achievement as measured by the state

mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Null Hypothesis 4: There are no statistically significant, research demonstrated, student variables that predict student mathematics achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Null Hypothesis 5: There are no statistically significant, research demonstrated, school variables that predict student language arts achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Null Hypothesis 6: There are no statistically significant, research demonstrated, school variables that predict student mathematics achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Sample Population/Data Source

All student data explored in this study pertained to students in Grade 8 enrolled in four of the five middle schools located in a suburban/urban central New Jersey community during the 2008-2009 academic school year. The student sample population for this study was 670 Grade 8 students. Data from students who met the following criteria were included in the study (a) general education program; (b) received a valid score on both sections of the 2009 NJ ASK language arts and mathematics tests; (c) received a valid score on both sections of the FAT pretest and posttest assessments in language arts and mathematics during the 2009 school year; and (d) do not qualify for the district's English language learners (ELL) program.

According to the New Jersey Department of the Treasury Local Government Budget Review (2001) the DE District in which the schools are located is one of the Top 10 largest in New Jersey comprised of 16 elementary, 5 middle, and 3 high schools. In 2008 approximately 100,000 people resided in the township (U.S. Census Bureau). Of those residents, 13,477 were children and adolescents that currently attend the DE District's 24 schools (TeacherPortal, 2009). Information obtained by the American FactFinder of the U.S. Census Bureau indicates 3.2% of township families fall below the poverty line (2005-2007). The Census Bureau verifies if a family is in poverty by calculating "if the family's total income is less than the family's threshold" (p.1) which would then determine if "that family and every individual in it is considered in poverty" (p.1).

Instrumentation

My intention was to determine if a significant relationship existed between student and school variables found in the extant literature to influence student achievement and aggregate district student NJ ASK scores in Grade 8 language arts and mathematics. Instrumentation for the study consisted of proficiency levels on scores for the state test, NJ ASK in Grade 8 as well as internal district formative pre and post FAT assessments results. Instrumentation is discussed in further detail below.

New Jersey Assessment of Skills and Knowledge Standardized Test

The assessment currently used by New Jersey is the NJ ASK which is administered in Grades 3-8 for language arts, mathematics, and science content areas to monitor the state`s progression toward reaching AYP targets. The NJDOE (2006b) maintains the purpose of the NJ ASK8 is to adequately and sufficiently verify that students are on par to pass the state mandated grade 11 assessment necessary for high school graduation. The NJ ASK6 and 7 are used as interim assessments to help the educators of each school district monitor progress toward obtaining the 2014 goal of one-hundred percent proficiency for all students. Ensuring New Jersey students succeed and thrive in a highly competitive global environment is the paramount objective of the NJ DOE.

The NJ ASK 2008 Technical Report confirmed:

> The NJ ASK Language Arts Literacy, Mathematics, and Science scores at grade 5-8 are reported as scale scores, with score ranges as follows:
>
> *(a)* Partially Proficient 100-199
>
> *(b)* Proficient 200-249
>
> *(c)* Advanced Proficient 250-300
>
> The scores of students who are included in the Partially Proficient level are considered to be below the state minimum of proficiency and those students may be most in need of instructional support. (p. 3)

Reliability

"The New Jersey Department of Education is required by federal law to ensure that the instruments it uses to measure student achievement for school accountability provide reliable results" (NJDOE, 2009, p. 116). The NJ ASK assessments were created under the auspices of Classical Test Theory (CTT).

The foundation of CTT is built upon the ideals that a total test score is comprised of multiple items. The CTT approach assumes "that the raw score (*X*) obtained by any one individual is made up of a true component (*T*) and a random error (*E*) component: X=T+E" (Kline, 2005, p. 91). Taking a person's mean score on the same test providing they had an infinite number of testing sessions would be the only manner in which one may obtain a person's true score. Since that is an impossibility, the central aspect of CTT is *T*, although this number is merely hypothetical (Kline, 2005).

Because high-stakes decisions are made often using solely a student's test scores (Tienken, 2008a, b), it is important to discuss the standard error of measurement (SEM) associated with these assessments. When considering the conditional SEM, Tienken (2008b, citing Harville, 1991) summarized that it is "an estimate of the amount of error or lack of precision one must consider when interpreting a test score" (p.37). He continued by stating that "the SEM describes how far the reported results may differ from a student's true score" (p.37). To clarify on the issues of reliability and SEM, the NJDOE (2009) published in the following in the NJ ASK Technical Report:

> Although the conceptualization of reliability and SEM is relatively straightforward, issues underlying the estimation of reliability are not. Reliability can be estimated via the correlation of scores on parallel forms or from test-retest data, or it can be estimated from a single test administration using any one of a variety of techniques (e.g., Brown, 1910; Cronbach, 1951; Kuder & Richardson, 1937). A very popular technique for estimating reliability from a single test administration is Cronbach's coefficient alpha. (p.117)

The NJ ASK Technical Report provided the following explanation regarding test metrics and units of analysis:

> The NJ ASK quantifies student achievement on three different metrics: number correct raw score, IRT scale, and performance score. While it is the knowledge and skills of individual students that are measured, student scores are aggregated and disaggregated into various units (e.g., school by grade, student group by grade, school, district, and state). Measurement error specific to each metric and each unit of analysis is taken into account when results are reported and accountability decisions are made. It is the responsibility of test developers to maximize reliability and minimize error by (1) identifying likely sources of error; (2) controlling the conditions of error; (3) estimating the size of error and/or level of reliability; and (4) reporting the estimates by metric and unit of analysis. (p. 117)

How closely related a set of items are as a group is known as internal consistency. In order to provide a unique estimate of the reliability for a given test. Cronbach's alpha was the reliability technique used for the NJ ASK. A statistical tool known as reliability coefficient is used to determine the extent to which a measure is reliable (Reinard, 2006). Reinard (2006) explained that:

> Reliability coefficients should be as close to 1.00 as possible, but interpretations often are based on guidelines such as the following:
>
> .90 and above: highly reliable
>
> .80-.89: good reliability
>
> .70-.79: fair reliability

> .60-.69: marginal reliability
>
> under .60: unacceptable reliability. (p.121)

Tienken (2008b, citing Frisbie, 1988, Rudner & Schafer, 2001) also reported "a reliability estimated of at least .85 out of a possible 1.00 should be used when an education leader makes high-stakes decisions about students, although an argument can be made for a minimum of .90 - .95" (p.36). Tienken emphasized the importance of reliability awareness necessary for those responsible for making high-stakes decisions. Because test results must be reported publically by all state education agencies (SEA) according to the NCLB Act regulations, there is some variation as to how this information is presented. Some SEA release proficiency percentages, while "some states go further and provide results for the specific sub-domains/content clusters of each full test (e.g., vocabulary, interpreting text, or narrative writing portions of the language arts section)" (Tienken, 2008b, p.36). The Language Arts Literacy (LAL) NJ ASK coefficient alpha score for Grade 8 was reported as 0.90 with the SEM of 3.17. The Math NJ ASK coefficient alpha score for Grade 8 was reported as 0.92 with the SEM of 3.25 (Tables 9.1.1 from NJ ASK Technical Report, 2009, p. 119). In a content analysis regarding characteristics of state assessment results, Tienken (2008b) concluded:

> Education leaders in one state may have access to a deeper understanding of the extent to which various technical factors, such as SEM, may influence usability of assessment results to make high stakes decisions about students, whereas educators in a neighboring state may not be aware of the potential issue. (p. 38)

It is important to clarify that although the coefficient alpha scores for language arts and mathematics were both in the 90s those percentages represented the coefficient alphas and SEM for the entire subject contents of the tests. However, when content sections were broken down into content clusters for each portion of the test, a much different image emerged regarding coefficient alphas. The table below includes the coefficient alphas for NJ ASK8 language arts and mathematics clusters.

Table 1

Coefficient Alphas of NJ ASK8 Content Clusters

Content Areas & Clusters		Alphas
LAL		0.90
Writing		0.67
Reading		0.89
	Working with Text	0.83
	Analyzing Text	0.78
Math		0.92
	Number & Numerical Operations	0.77
	Geometry & Measurements	0.69
	Patterns & Algebra	0.77
	Data Analysis, Probability, & Discrete Mathematics	0.71
	Problem Solving	0.88

Notably, once content clusters for language arts are analyzed individually the coefficient alpha percentages clearly decrease from the 90s and in many areas below the acceptable reliability estimate of at least 0.85. Most concerning is the recorded 0.67 in the writing content cluster. According to Reinard (2006) a 0.67 coefficient alpha falls in the range of marginally reliable, yet district leaders often use this score of a student to make course tracking decisions for high school entrance. For mathematics, the geometry and measurement cluster also meets the qualifications of a marginally reliable measure at 0.69. The reliability of the test scores is based of the statewide population and may not be representative of the DE District students. Thus, data used in my study may be different.

Validity

According to Baker and Linn (2002) "two questions are central in the evaluation of content aspects of validity. Is the definition of the content domain to be assessed adequate and appropriate? Does the test provide an adequate representation of the content domain the test is intended to measure?" (p. 6). The NJDOE claimed that the answers to these two guiding questions can be located in the NJ ASK Technical Report (2009). The NJDOE maintained that the appropriateness of content of the NJ ASK assessments acknowledging that all tests grades 3-8 were in alignment with the NJCCCS. The NJCCCS are the framework in which educators follow that identifies what all students should know and be able to do within the designated grade level.

Regarding the adequacy of content representation, the NJDOE (2009) acknowledged it "is critically important because the tests must provide an indication of student progress toward achieving the knowledge and skills identified in the CCCS, and the tests must fulfill the requirements under NCLB" (p. 143). The NJDOE assured that through the use of a test blueprint and a responsible test construction process this measure is properly adhered to. The NJDOE explained in the NJ ASK Technical Report that:

> New Jersey performance standards, as well as the CCCS, are taken into consideration in the writing of multiple-choice and constructed response items

> and constructed-response rubric development. Each test must align with and proportionally represent the sub domains of the test blueprint. (p. 143)

Construct validity is defined as the:

> Validity of a test or a measurement tool that is established by demonstrating its ability to identify or measure the variables or constructs that it proposes to identify or measure. The judgment is based on the accumulation of correlations from numerous studies using the instrument being evaluated. (Mosby's Medical Dictionary, 2009)

In a section titled "Construct Validity" of the NJ ASK Technical Report (2009) the NJDOE purported:

> Because the NJ ASK testing program assesses student performance in several content areas using a variety of testing methods, it is important to study the pattern of relationships among the content areas and testing methods. Therefore, this section addresses evidence based on responses and internal structure. One method for studying patterns of relationships to provide evidence supporting the inferences made from test scores is the multi-trait matrix. Tables 7.3.1 through 7.3.4 summarize Pearson correlation coefficients among test content domains and clusters by grade level. The correlations between clusters within a content area were generally found to be higher than the correlations between clusters across the content areas. (p. 144)

Construct validity is also influenced by the way test results are used. Construct validity issues are presented when a single standardized test score is used to make a judgment about a student's overall ability in a subject area. "The traditional view of validity as three distinct categories, construct, content, and criterion is ill-suited to explain the potential negative social and education consequences of test-score misinterpretation" (Tienken & Rodriguez, 2010, p.164). Messick (1995) analyzed the validity of integrated criteria and content, as a result, the construct validity framework consequences both intended and unintended were revealed. Rather than place the social and educational consequences of test score interpretation into a separate validity category, Messick classified them as an aspect of construct validity. "The integrated view of construct validity allows education administration and policymakers to consider social and education consequences in the validity discussion" (Tienken & Rodriguez, 2010, p.164).

The Commercially Produced Formative Assessment Tool (FAT)

The following information regarding FAT was obtained through The Company website as well as extensive conversations and email correspondings with the DE District FAT liaisons and The Company representatives.

As of 2009, FAT is available to school districts in California, New Jersey, and Texas. FAT is a web-based technology formative assessment tool produced by The Company. FAT is designed to help diagnose the academic strengths and weaknesses of individual students in language arts and mathematics, two content areas that must annually meet AYP targets. Assessments delivered by FAT are in alignment with the NJCCCS as well as with the criteria set forth by Measurement Incorporate (MI), the educational company that develops and scores the NJ ASK. It is anticipated that individual student information provided to educators via FAT can assist in increasing student achievement levels on the NJ ASK.

A computer generated pre assessment is produced by FAT for language arts and mathematics in the fall at the commencement of the school year. After students are administered the test, data results are immediately available for educators to analyze. This early pre assessment allows sufficient amount of time for educators to differentiate instruction and focus on existing achievement gaps that may not have been known prior to testing. Data produced by FAT is available on multiple levels for language arts and mathematics: district, school, grade, teacher, and student. Detailed results accentuate individual student's achievements and misunderstandings which allow educators to focus exclusively on skills that students need to improve upon while moving along with ones that have already been mastered. Individual online interim assessments can be created and assigned to students at the discretion of the educator. These assessments can be developed using either FAT's provided question banks or by uploading unique teacher generated questions which vary from the pre assessment questions.

Over the course of the school year educators are encouraged to assess their students at different intervals using a variety of classroom practices and formative assessment techniques. Prior to the state administration of the NJ ASK in April, a FAT post assessment is administered and these results are compared with the students` pre assessment results. Educators learn of additional areas of support students may need in order to successful pass the impending NJ ASK.

During the 2007-2008 school year, FAT was introduced and offered to all New Jersey school districts free of charge for 5 years. After the 5 year trial period expires in the 2011-2012 school year, district administrators must decide whether to purchase the product from The Company for continued use. District administrators were encouraged to take advantage of the free web-based formative assessment system to monitor student achievement. District administrators could decide at any point during the 5 year time period to implement FAT. The DE District did not opt to use FAT during the 2007-2008 school year, but did implement it at the start of the 2008-2009 school year. The Company fully supports districts that use FAT and continually provides professional development opportunities to ensure optimum product use.

Data Collection

Permission was granted to me as the researcher to use all the requested sources of information by the DE District's Superintendent of Schools and Assistant Superintendent of Curriculum and Instruction. All data was collected by the DE District's employees responsible for handling that particular data source. For example, the NJ ASK test scores provided by the state reports were converted into an Excel spreadsheet and given to me by the District's Test Coordinator. The free and reduced lunch report, attendance records, and ASI class rosters were provided by the District's Data Analyst. FAT pre and post assessment results were collected personally by me via the FAT summary report feature. All reports were coded to guarantee confidentially. Each coded student identifier represented a record. Each complete report contained the following data unique to each record: NJ ASK test scores (language arts and math), FAT pre assessment scores (language arts and math), FAT post assessment scores (language arts and math), free and reduced lunch identification, attendance record, and ASI eligibility. Incomplete records void of a least one component of data were excluded from the study.

According to the DE District's NJDOE 2008-2009 School Report Cards, the following percentages of faculty/administrators held either a MA or MS degree, the breakdown is as follows: School A had 19.4%; School B had 37.5%; School C had 39.4%, School D had 33.8%; and School E 30.2%. It is important to explain that faculty and administrators include principals, vice principals, school nurses, media specialists, guidance counselors, social workers, Gifted and Talented staff, among many other positions . Therefore, it is important to identify and separate language arts and mathematic teachers that hold master's degrees specifically in the content area that they provided instruction in. In the DE District there is very little variance in the middle school teachers that held a master's degree in their subject content (language arts or mathematics) during the 2008-2009 school year. The break down per school is outlined in Table 2.

Table 2

Language Arts and Mathematics Teachers Holding Content Specific Master's Degrees by School

School	N= English/Reading	N= Mathematics
A	0	3
B	1	0
C	2	2
D	4	1
E	0	1

When looking specifically at Grade 8 teachers the numbers are similar to those presented in Table 2.

Analysis Construct

Figure 2 is a production function theory diagram that guided the data analysis. Note that, in comparison to the diagram found in Chapter II, Academic Support Instruction (ASI) has been added to the production function model and teachers that held master's degree in subject content area has been removed due to a lack of overall variance in teacher degree status in the district.

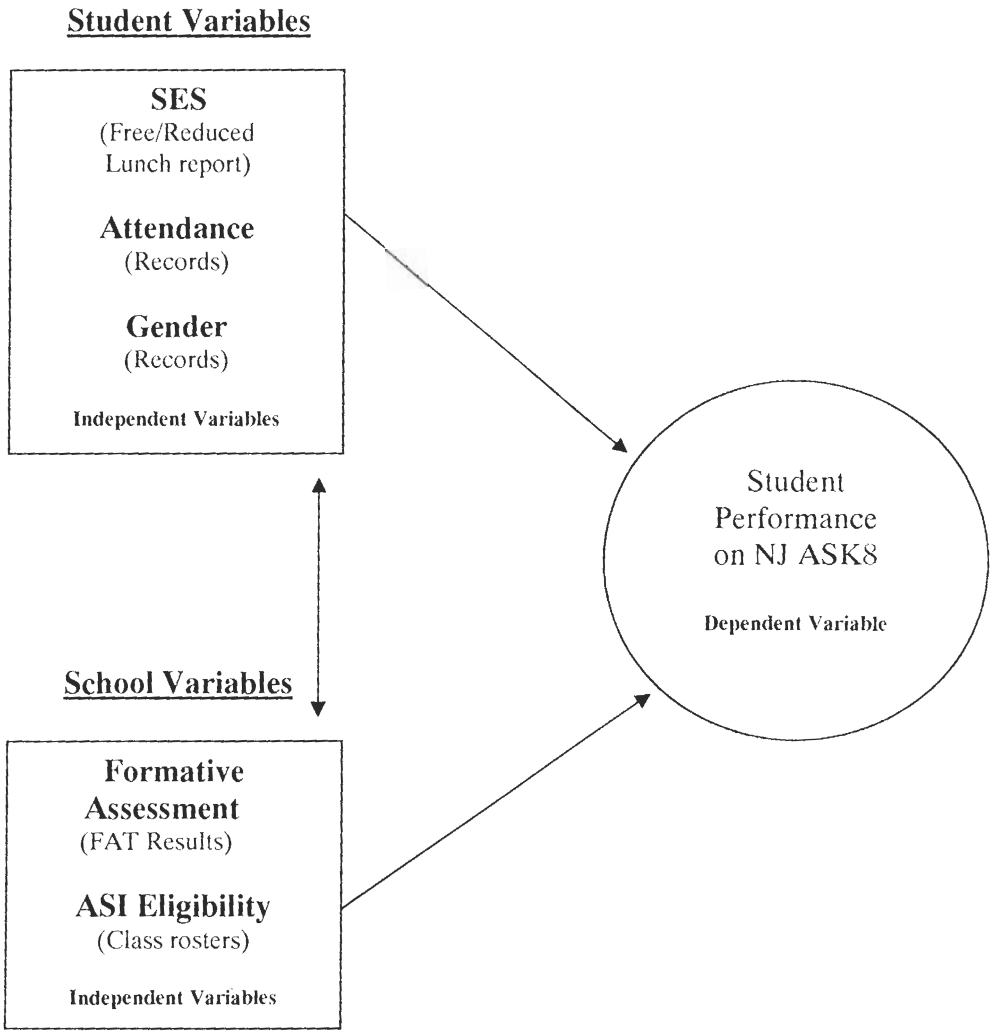


Figure 2. Input/output framework modified.

Due to the limited number of content specific degrees held by middle school teachers, that variable has been removed from my model and replaced with ASI eligibility. In an effort to "specialize instruction based on the student's individual needs" (2008-2009 Middle School Program of Studies Guide) the DE District provides ASI in both language arts and mathematics in Grades 6-8. The DE District's Middle School Program of Studies Guide reported for ASI Mathematics that:

> Students will be required to enroll in this course if they scored below the designated levels of proficiency in the state's NJ ASK – 5, 6, or 7. Input from classroom teachers, guidance personnel, and building principals is also considered before final placement in this program is recommended. The content of this course includes topics from the appropriate grade-level mathematics curriculum as well as specialized instruction based on the student's individual needs. (p.11)

In ASI Language Arts it was reported that:

> Students will be required to enroll in this course if they scored below the designated levels of proficiency in the state's NJ ASK – 5, 6, or 7. Input from classroom teachers, guidance personnel, and building principals is also considered before final placement in this program is recommended. Courses Include: ASI Reading/Writing Workshop (Grade 6) and ASI Language Arts Workshop (Grades 7 & 8). The content of these courses includes topics from the regular grade-appropriate Reading/Writing and Language Arts courses as well as specialized instruction based on the student's individual needs. (p.11)

ASI Mathematics and ASI Language Arts courses are conducted during the regular school day and replace student's mandatory mathematics and language arts classes. ASI is not an after school or pull-out program. The inclusion of the ASI eligibility variable can provide useful information in regards to predictors that influence student achievement. School administrators exert direct control over entrance criteria, teachers assigned to teach the program, and curriculum. It is a variable that, if statistically significantly related to student achievement in some way, school administrators can mutate in an attempt to improve student achievement in the future. It is valuable for community stakeholders to ascertain if individualized instruction courses similar to the DE District's ASI program truly improve student achievement.

Data Analysis

My initial intentions were to include the data from the five middle schools in the DE District in my study. However, after running the data for the fifth school, School E, I identified that major regression violations had occurred as a result of the small sample size (98) in comparison with the number of independent variables (8). Therefore, School E was removed from the study; however the characteristics regarding the school remain within this chapter and are included along with the other schools.

I analyzed the data from the four schools separately instead of aggregating the scores. Separate analyses were conducted for two reasons: (a) The district leadership interprets the test results on the individual school level and makes decisions for each school based on that school's results. Resources are allocated based on the individual schools' output, not the aggregate of all the district's middle schools. (b) Due to significant differences in student scale scores in language arts and mathematics on the NJ ASK8 between schools, each school was analyzed individually. Tables 3-9 outline the demographic characteristics as well as scale score differences and proficiency

Table 5

Demographic Characteristics of Grade 8 Students by School

Ethnicity Breakdown									Total
School	Total	White	Black	Asian	Pacific Islander	Hispanic	Amer. Indian	Other	Econ. Dis.
A	227	97	49	33	1	47	0	0	80
B	229	165	19	22	0	22	1	0	31
C	200	70	27	32	0	70	1	0	89
D	252	97	29	99	0	27	0	0	58
E	172	96	20	10	0	46	0	0	41

Table 6

NJ ASK 2009 General Education Scale Score Mean for LA & Math

School	Grade 6 LA	Grade 6 Math	Grade 7 LA	Grade 7 Math	Grade 8 LA	Grade 8 Math
A						
General Ed.	216.2	239.7	215.3	228.5	224.4	235.2
Econ. Disadv.	205.0	226.8	203.3	216.9	215.4	220.6
Non-Econ. Disadv.	216.6	239.0	214.8	225.7	223.1	235.1
B						
General Ed.	225.1	241.2	227.4	237.3	230.5	240.5
Econ. Disadv.	217.2	224.7	208.5	217.4	214.3	216.5
Non-Econ. Disadv.	222.4	240.0	224.4	233.2	228.4	236.3

School	Grade 6 LA	Grade 6 Math	Grade 7 LA	Grade 7 Math	Grade 8 LA	Grade 8 Math
C						
General Ed.	216.8	232.0	220.2	229.4	223.2	234.3
Econ. Disadv.	201.5	208.2	203.0	210.1	211.8	213.1
Non-Econ. Disadv.	216.6	233.3	221.0	229.2	220.1	230.1
D						
General Ed.	221.8	254.5	228.9	247.2	230.4	249.4
Econ. Disadv.	205.0	228.0	210.2	226.9	217.5	223.5
Non-Econ. Disadv.	220.2	250.7	227.3	244.2	228.5	244.8
E						
General Ed.	213.8	231.5	222.8	240.8	226.9	248
Econ. Disadv.	198.7	213.6	209.3	226.6	218.5	230.1
Non-Econ. Disadv.	216.9	234.2	218.4	236.5	226.2	247.1

Table 7

NJ ASK 2009 Grade 6 General Education Proficiency Percentages for LA & Math

	Language Arts			Mathematics		
School	PP	P	AP	PP	P	AP
A	19.8	72	8.2	12	43.2	44.8
B	11.3	71.6	17	13.9	41.2	44.8
C	20.6	70.4	9	20.6	47.6	31.7
D	13.9	72.1	13.9	4.2	40	55.8
E	29.8	58.7	11.6	17.6	48.7	33.6

Note. PP = Partially Proficient (students did not pass the test); P = Proficient; AP = Advanced Proficient.

Table 8

NJ ASK 2009 Grade 7 General Education Proficiency Percentages for LA & Math

	Language Arts			Mathematics		
School	PP	P	AP	PP	P	AP
A	27.1	58.6	14.4	19.3	47.0	33.7
B	13.1	58.6	28.3	15.2	45	39.8
C	19.2	63.8	17	21.4	46.9	31.7
D	12.8	61.6	25.6	7.6	37.8	54.7
E	16.3	63.6	20.2	14.7	36.4	48.8

Note. PP = Partially Proficient (students did not pass the test); P = Proficient; AP = Advanced Proficient.

Table 9

NJ ASK 2009 Grade 8 General Education Proficiency Percentages for LA & Math

	Language Arts			Mathematics		
School	PP	P	AP	PP	P	AP
A	9.6	82.3	8.1	14.1	48	37.9
B	4	76.3	19.7	13.6	42.9	43.4
C	8.9	80.4	10.8	17.7	44.9	37.3
D	4.6	76	19.4	9.2	36.9	53.9
E	4.6	85	10.5	5.3	42.4	52.3

Note. PP = Partially Proficient (students did not pass the test); P = Proficient; AP = Advanced Proficient.

Since my study included four middle schools within the same district and both language arts and mathematics were researched, nine multiple regression models were presented. All collected data was inputted in SPSS version 16. Through the use of multiple regression analysis, the predictor variables (i.e., student and school variables) were inputted as the independent variables whereas the NJ ASK8 scores were inputted as the dependent variable, which is a scale level, dependent variable. According to Leech, Morgan, and Barrett (2008):

> The assumptions for multiple regression include the following: that the relationship between each of the predictor variables and the dependent variable is linear and that the error, or residual, is normally distributed and uncorrelated with the predictors. (p. 95)

Since the independent variables I worked with were not all continuous (i.e., free/reduced lunch eligibility, attendance policy, and ASI eligibility), I used dummy coding for categorical variables. A variable is considered dichotomous when it takes on only two values. In the case of these three variables the values could only be "yes" or "no". For example, "yes" the student is eligible for free/reduced lunch or "no" the student is not eligible for free/reduced lunch. The following recoding was used for the student variable of SES: 1 = eligible for free/reduced lunch, 0 = not eligible; for the student variable of attendance: 1= student exceed district policy of 16 absences, 0= student did not exceed 16 absences; for the school variable of ASI eligibility: 1= eligible for ASI services, and 0= not eligible. Gender was coded as 0=male, 1= non-male.

As mentioned previously, the NJ ASK scores range from any of three levels, Partially Proficient (PP) 100-199, Proficient (P) 200-249, and Advanced Proficient (AP) 250-300. FAT pretest and posttest scores range from any of four proficiency levels, Below Basic 0-54, Basic 55-70, Proficient 71-85, and Advanced 86-100. FAT representatives maintained that both the Basic and the Proficient score ranges are equal to the Proficient scale on the NJ ASK, while the FAT scale of Advanced correlates with the

Advanced Proficient level on the NJ ASK. FAT LA and Math pretest and posttest variables were coded as: 1= proficient, and 0= not proficient.

Multicollinearity "happens when two or more predictors contain much of the same information" (Leech, Morgan, & Barrett, 2008, p. 94). In an effort to avoid this from happening while adhering to the multiple regression assumption "that the relationship between each of the predictor variables and the dependent variables is linear and that the error, or residual, is normally distributed and uncorrelated with the predictors" (p.95), I first checked the correlations among the independent variables. Multicollinearity was also detected by computing the variance inflation factor (VIF) from the data loaded into the regression models.

By using the Enter method of the SPSS program (also known as simultaneous regression), all appropriate student and school variables were entered at the same time. From this I was able to ascertain which predictors contributed statistically significantly to the multiple regressions. After which I "create[d] a scatterplot matrix to check the assumption of linear relationships of each predictor with the dependent variable and a scatterplot between the predictive equation and the residual to check for the assumption that these are uncorrelated" (Leech, Morgan, & Barrett, 2008, p. 95). After the correlations were generated I checked for high correlations among predictors, and if necessary, eliminated variables that displayed multicollinearity.

I ran two multiple regression analyses for each school, one for language arts and one for math, and therefore generated a total of eight model summaries. It was important to analyze the summaries and determine the values of R, which indicated the multiple correlation coefficients and the adjusted R squares, which identified the percentage of the variance in the dependent variable that was predicated from the independent variables. ANOVA tables reported the F statistics and determined whether or not the combination of the student and school predictor variables statistically significantly predicted student achievement on the NJ ASK in language arts and mathematics. Also included were nine coefficients tables that produced valuable information (one school included a Corrected Language Arts Model). Here the standardized beta coefficients were revealed. From the *t* value and the p value, I was able to determine if one specific variable statistically significantly contributed to the prediction equation for NJ ASK scores from all the independent variables. However, I painted a clearer picture and uncovered more information than what was revealed by reporting just beta weights.

According to Thompson (2006) a regression structure coefficient is "the bivariate Pearson *r* of a measured predictor with the latent Yhat scores" (p.240). Yhat represents the latent variable. Thompson (2006) recommends that researchers interpret (a) the beta weights and the structure coefficients or (b) beta weights and the bivariate correlations of the predictors with the Y variable, but never just the beta weights. The time when the researcher should interpret only beta weights is in cases when there is only one predictor (Courville & Thompson, 2001). When predictor variables are correlated with each other, as they often are, results from regression analyses can be misinterpreted. Whitaker (1997), stated:

> The unwary research might be tempted to regard the predictor variable with the largest absolute value as the greatest predictor.... It is possible to have a predictor variable with the greatest predictive potential lose credit to two (or more) other predicators whose predictive area overlaps that of the first predictor. The first

> predictor is given no credit for predictive potential and could have a beta weight of zero. In this instance, it is important to have information about the true predictive potential of that variable, information that can be easily gained by examining each predictor variable's structure coefficient. (p.7)

Structure coefficients are not restricted by statistical significance and they are not affected by collinearity. In fact, statistical significance and collinearity are not concerns when interpreting structure coefficients. Computing and then analyzing structure coefficients is an appropriate way to paint a more complete picture of the regression results and help to uncover potentially important predictors that would not be given credit if the researcher analyzed only beta weights.

Thompson (2006) reported that some researchers presented objections to reporting structure coefficients in regression. The main objection revolves around structure coefficients not being affected by collinearity among the predictor variables, whereas the beta weights are affected by correlations among predictors. This perceived insensitivity of structure coefficients on the part of some is misplaced. Thompson (2006) wrote that although it is true that "...beta weights are context-specific to a particular set of measured predictor variables" (p241), the insensitivity of structure coefficients to changing contexts should not be viewed as a weakness. Thompson stated:

> Because science is about the business of generalizing relationships across participants, across variables and measures of variables, and across time, in some respects it is desirable that structure coefficients are not impacted by collinearity. This insensitivity honors the reality in which measured predictors variables are

> correlated, and structure coefficients are unaffected by this colliearity, which is instead duly considered when computing beta weights. (p.242)

Chapter IV

ANALYSIS OF THE DATA

Introduction

The purpose of this quantitative study was to investigate the influence of select student and school variables, found in the extant literature, on student achievement as measured by Grade 8 NJ ASK in language arts (LA) and mathematics. By placing the focus on multiple student and school variables that have a statistically significant relationship to student achievement, this study produced research-based evidence that could assist stakeholders in middle class and lower middle class public schools in New Jersey regarding reform initiatives.

In order to determine which student and school variables demonstrated a statistically significant relationship to student achievement, I used the simultaneous multiple regression model to analyze the data. This strategy is used when the researcher has no logical or theoretical structure of the data. This method is typically used to explore and maximize prediction (Pedhazur, 1997). Scatter diagrams of residuals and normal probability plots of residuals were conducted to test assumptions.

Analysis Strategy

I analyzed the data from four middle schools in the district separately instead of aggregating the scores. Separate analyses were conducted for two reasons: (a) The district leadership interprets the test results on the individual school level and makes decisions for each school based on that school's results. Resources are allocated based on the individual schools' output, not the aggregate of all the district's middle schools. (b) Due to significant differences in student demographics and student achievement characteristics in language arts and mathematics on the NJ ASK8 between schools, each school was analyzed individually.

The F static was used to determine whether the regression models were statistically significant. The coefficient of determination, R^2, was interpreted as the proportion of the variance in the dependent variables (NJ ASK8 LA and Math scores) that is predictable from the independent variables (student variables and school variables). Adjusted R^2 (adj R^2), is a modification of R^2 that served the same purpose. Adjusted R^2 is generally considered to be a more accurate measure that adjusts for the number of explanatory terms in a model, and therefore was used in my regression analyses. Reported in each regression model are the number of values in the final calculation of a statistic that are free to vary; this term is referred to as the degrees of freedom (df). The standardized beta coefficient, most commonly referred to as the beta, was used to compare the strength of the effect of each independent variable (student and school) on the dependent variables (NJ ASK8 LA and Math). The *t* statistic is used to determine whether or not the effect of the independent variable on the dependent variable is significant.

I examined the tolerance values for the predictors in each model as a check for multicollinearity. Multicollinearity is a condition that can be problematic and lead to misinterpretation of the results. It occurs when there are high correlations among some of the predictor variables in the model. When two or more predictors contain the same information, there is multicollinearity. Since prediction credit in a regression model cannot be shared among variables it is important to test for multicollinearity. If the tolerance value is low ($<1-R^2$), then there could be a problem with multicollinearity and variables must be eliminated or combined.

Structure Coefficients

According to Thompson (2006) a regression structure coefficient is "the bivariate Pearson *r* of a measured predictor with the latent Yhat scores" (p.240). Yhat represents the latent variable. Thompson (2006) recommends that researchers interpret (a) the beta weights and the structure coefficients or (b) beta weights and the bivariate correlations of the predictors with the Y variable, but never just the beta weights. The time when the researcher should interpret only beta weights is in cases when there is only one predictor (Courville & Thompson, 2001). When predictor variables are correlated with each other, as they often are, results from regression analyses can be misinterpreted. Whitaker (1997), stated:

> The unwary research might be tempted to regard the predictor variable with the largest absolute value as the greatest predictor.... It is possible to have a predictor variable with the greatest predictive potential lose credit to two (or more) other predicators whose predictive area overlaps that of the first predictor. The first predictor is given no credit for predictive potential and could have a beta weight of zero. In this instance, it is important to have information about the true predictive potential of that variable, information that can be easily gained by examining each predictor variable's structure coefficient. (p.7)

Structure coefficients are not restricted by statistical significance and they are not affected by collinearity. In fact, statistical significance and collinearity are not concerns when interpreting structure coefficients. Computing and then analyzing structure coefficients is an appropriate way to paint a more complete picture of the regression results and help to uncover potentially important predictors that would not be given credit if the researcher analyzed only beta weights.

Thompson (2006) reported that some researchers presented objections to reporting structure coefficients in regression. The main objection revolves around structure coefficients not being affected by collinearity among the predictor variables, whereas the beta weights are affected by correlations among predictors. This perceived insensitivity of structure coefficients on the part of some is misplaced. Thompson (2006) wrote that although it is true that "...beta weights are context-specific to a particular set of measured predictor variables" (p241), the insensitivity of structure coefficients to changing contexts should not be viewed as a weakness. Thompson stated:

> Because science is about the business of generalizing relationships across participants, across variables and measures of variables, and across time, in some respects it is desirable that structure coefficients are not impacted by collinearity. This insensitivity honors the reality in which measured predictors variables are correlated, and structure coefficients are unaffected by this colliearity, which is instead duly considered when computing beta weights. (p.242)

Strong correlations among predictor variables can cause reductions in the standardized coefficients (beta) because predictive credit can only be given to one variable, even when the variance overlaps between two correlated predictors. Only one variable can receive the credit. Therefore, there are cases in which a predictor variable correlates strongly with the dependent variable, but is assigned a near-zero beta weight, thus indicating that the predictive weight of that variable is being sapped by another variable. The individual SPSS data outputs produced will be used to answer the following research questions for each of the four middle schools:

Research Question 1: What is the strength and direction of the relationship between the proficiency categorizations of students in Grade 8 measured by commercially produced standardized formative assessment in language arts and the proficiency categorizations of students in Grade 8 measured by the language arts portion of the NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Research Question 2: What is the strength and direction of the relationship between the proficiency categorizations of students in Grade 8 measured by commercially produced standardized formative assessment in mathematics and the proficiency categorizations of students in Grade 8 measured by the mathematics portion of the NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Research Question 3: What are the statistically significant student variables that explain the largest amount of variance in student language arts achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Research Question 4: What are the statistically significant student variables that explain the largest amount of variance in student mathematics achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Research Question 5: What are the statistically significant school variables that explain the largest amount of variance in student language arts achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Research Question 6: What are the statistically significant school variables that explain largest amount of variance in student mathematics achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Results

Language Arts and Mathematics Models for School A

First I created a matrix scatterplot to determine if the variables were related to each other in a linear fashion and found that to be true (see Figure 3). The data are arranged in columns because dichotomous variables are plotted according to data points. "Linearity would be violated if the data points bunch at the center of one column and at the ends of the other column" (Leech, Barrett, & Morgan, 2008, p. 103). The scatterplot results suggested that the assumption of linearity was not violated.

Language arts.

A multiple regression analysis for School A was performed between the dependent variable (NJ ASK8 LA scores) and the independent variables (gender, NJ ASK8 Math scores, attendance, SES, FAT LA and Math Pretest scores, and FAT LA and Math Posttest scores). I performed a descriptive analysis initially (see Table 10).

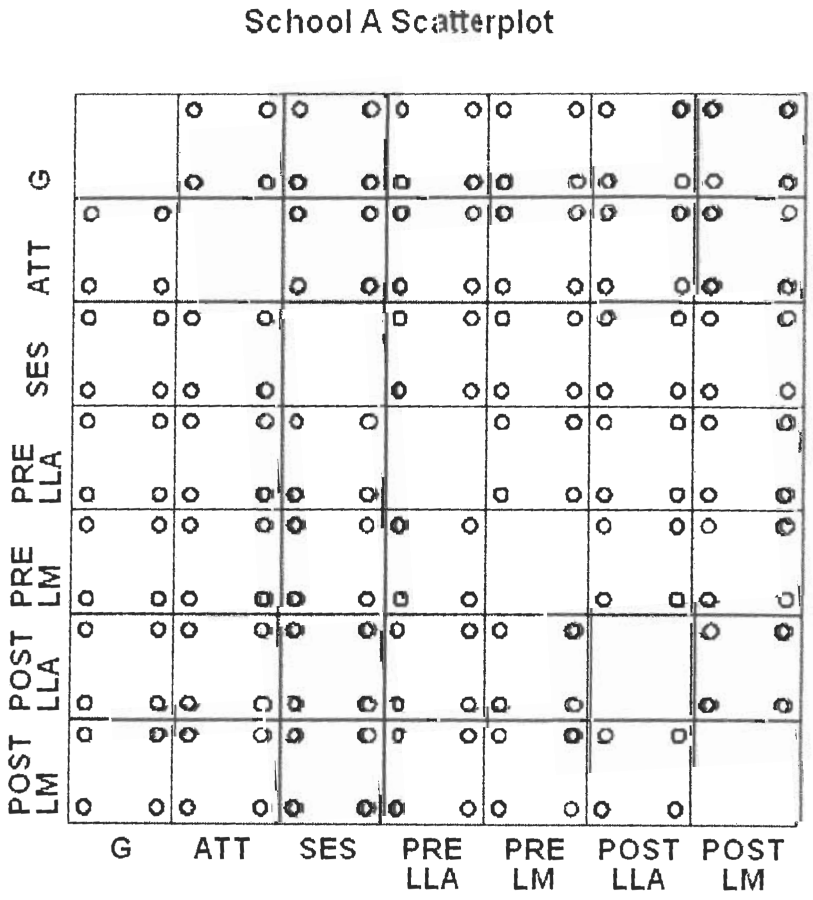


Figure 3. School A scatterplot.

Table 10

Descriptive Statistics for School A Language Arts

	Mean	Std. Deviation	N
ASK_LA	225.3235	18.33502	170
Gender 0=m	.5706	.49645	170
Attend	.2000	.40118	170
SES 0= Not	.3235	.46920	170
PRELA 0=Not Prof	.3294	.47139	170
PreM 0 = Not Prof	.2706	.44558	170
PostLA 0=Not Prf	.4824	.50116	170
PostM 0=Not Pr	.3000	.45961	170
ASK_M	235.7647	38.17901	170

Using the enter method, a statistically significant model emerged (*F*= 31.209, p=≤ .001, Adjusted R^2 .588). The model summary suggested that approximately 58% of the variance in student performance on the NJ ASK8 LA and coefficients revealed four statistically significant predictors (see Tables 11 & 12).

Table 11

Model Summary School A Language Arts

Model					Change Statistics				
	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.780^a	.608	.588	11.76189	.608	31.209	8	161	.000

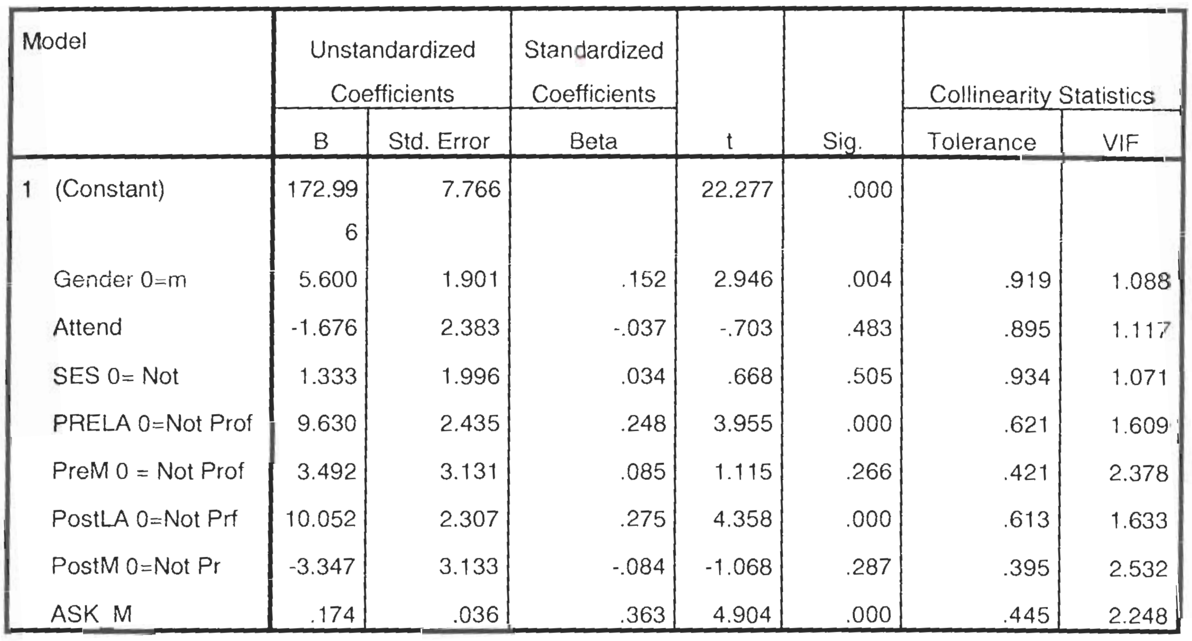
a. Predictors: (Constant), ASK_M, Gender 0=m, SES 0= Not, Attend , PRELA 0=Not Prof, PostLA 0=Not Prf, PreM 0 = Not Prof, PostM 0=Not Pr

b. Dependent Variable: ASK_LA

Multicollinearity was not an issue as the tolerance values for the predictors were not exceedingly low (<1-R^2).

Table 12

Coefficients for School A Language Arts



Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1 (Constant)	172.996	7.766		22.277	.000		
Gender 0=m	5.600	1.901	.152	2.946	.004	.919	1.088
Attend	-1.676	2.383	-.037	-.703	.483	.895	1.117
SES 0= Not	1.333	1.996	.034	.668	.505	.934	1.071
PRELA 0=Not Prof	9.630	2.435	.248	3.955	.000	.621	1.609
PreM 0 = Not Prof	3.492	3.131	.085	1.115	.266	.421	2.378
PostLA 0=Not Prf	10.052	2.307	.275	4.358	.000	.613	1.633
PostM 0=Not Pr	-3.347	3.133	-.084	-1.068	.287	.395	2.532
ASK_M	.174	.036	.363	4.904	.000	.445	2.248

a. Dependent Variable: ASK_LA

In this model, the variance explained (adj R^2) was .588, which indicated that approximately 58% of the variance in NJ ASK8 LA scores was explained by the variables in the model. Statistically significant predictors variables, their betas, and p values were as follows: (a) NJ ASK8 Math, .363, p<.001, (b) FAT LA Posttest, .275, p<.001, (c) FAT LA Pretest, .248, p<.001, and (d) gender, .152, p=.004.

Mathematics.

A second multiple regression analysis for School A was performed between the dependent variable (NJ ASK8 Math scores) and the independent variables (gender, NJ ASK8 LA scores, attendance, SES, FAT LA and Math Pretest scores, and FAT LA and Math Posttest scores). I performed a descriptive analysis initially (see Table 13).

Table 13

Descriptive Statistics for School A Mathematics

	Mean	Std. Deviation	N
ASK_M	235.7647	38.17901	170
Gender 0=m	.5706	.49645	170
Attend	.2000	.40118	170
SES 0= Not	.3235	.46920	170
PRELA 0=Not Prof	.3294	.47139	170
PreM 0 = Not Prof	.2706	.44558	170
PostLA 0=Not Prf	.4824	.50116	170
PostM 0=Not Pr	.3000	.45961	170
ASK_LA	225.3235	18.33502	170

Using the enter method, a statistically significant model emerged (*F* = 31.866, p < .001, Adjusted R^2 .594.) The model summary revealed that approximately 59% of the variance in student performance on the NJ ASK8 Math and the coefficients revealed two statistically significant predictors (see Tables 14 & 15).

Table 14

Model Summary School A Mathematics

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.783[a]	.613	.594	24.33653	.613	31.866	8	161	.000

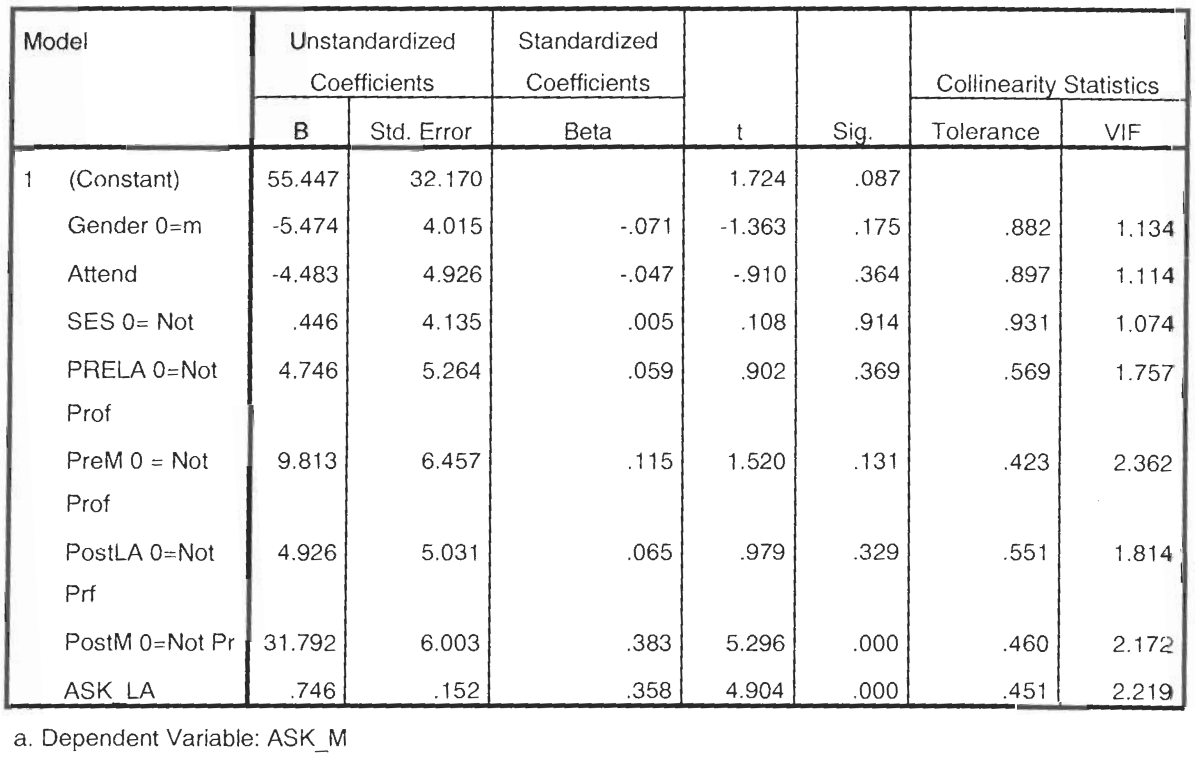
a. Predictors: (Constant), ASK_LA, SES 0= Not, Attend , Gender 0=m, PostM 0=Not Pr, PRELA 0=Not Prof, PostLA 0=Not Prf, PreM 0 = Not Prof

b. Dependent Variable: ASK_M

In this model, the variance explained (adj R^2) was .594, which indicated that approximately 59% of the variance in NJ ASK8 Math scores was explained by the variables in the model. Two variables were statistically significant predictors variables were, (a) FAT Math Posttest with a beta of .383, (p<.001) with a *t* value of 5.296, and (b) NJ ASK8 LA (p<.001) with a beta of .358 with a *t* value of 4.904. Multicollinearity was not an issue as the tolerance values for the predictors were not low (<1- R^2).

Table 15

Coefficients for School A Mathematics



Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	55.447	32.170		1.724	.087		
	Gender 0=m	-5.474	4.015	-.071	-1.363	.175	.882	1.134
	Attend	-4.483	4.926	-.047	-.910	.364	.897	1.114
	SES 0= Not	.446	4.135	.005	.108	.914	.931	1.074
	PRELA 0=Not Prof	4.746	5.264	.059	.902	.369	.569	1.757
	PreM 0 = Not Prof	9.813	6.457	.115	1.520	.131	.423	2.362
	PostLA 0=Not Prf	4.926	5.031	.065	.979	.329	.551	1.814
	PostM 0=Not Pr	31.792	6.003	.383	5.296	.000	.460	2.172
	ASK_LA	.746	.152	.358	4.904	.000	.451	2.219

a. Dependent Variable: ASK_M

Research Questions and Answers for School A

Research Question 1: What is the strength and direction of the relationship between the proficiency categorizations of students in Grade 8 measured by commercially produced standardized formative assessment in language arts and the proficiency categorizations of students in Grade 8 measured by the language arts portion of the NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Null Hypothesis 1: No statistically significant relationship exists between a commercially produced standardized formative assessment tool proficiency categorization and students' language arts proficiency categorization on the NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Answer: The null hypothesis is rejected. The commercially prepared FAT LA Pretest and FAT LA Posttest were statistically significant predictors of student achievement on the NJ ASK8 LA test. The FAT LA Posttest had a beta of .275, (p<.001) with a *t* value 4.358, (b) FAT LA Pretest, .248, (p<.001) with a *t* value of 3.955.

It should be noted that although NJ ASK8 Math was not the subject of Research Question 1, it was reported as the best predictor of NJ ASK8 LA achievement. However, the predictive power of the NJ ASK8 Math is suspect. It stands to reason that the true relationship comes from a student's language arts skills that influence his/her mathematics scores because of the amount of reading necessary to engage the NJ ASK8 mathematics test. The seemingly apparent influence of math achievement on language arts achievement might be a false finding and should be interpreted with caution.

As mentioned in Chapter III, beta weights present only part of the story when predictor variables are correlated. Thus, I computed structure coefficients. The structure coefficients confirmed that the FAT Posttest LA (r_s=.806) and FAT Pretest LA (r_s=.771) tests result variables were contributing to the prediction of achievement on the NJ ASK8 LA.

The results suggest that the FAT LA Pretest results have almost the same predictive power in terms of betas and structure coefficients as the FAT LA Posttest.

Research Question 2: What is the strength and direction of the relationship between the proficiency categorizations of students in Grade 8 measured by commercially produced standardized formative assessment in mathematics and the proficiency categorizations of students in Grade 8 measured by the mathematics portion of the NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Null Hypothesis 2: No statistically significant relationship exists between a commercially produced standardized formative assessment tool proficiency categorization and students' mathematics proficiency categorization on the NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Answer: The null hypothesis is rejected. In this model, the variance explained (adj R^2) was .594, which indicated that approximately 59% of the variance in NJ ASK8 Math scores by the variables in the model. The one statistically significant predictor variable was: (a) FAT Math Posttest with a beta of, .383, ($p<.001$) and a *t* value of 5.296.

Next, I computed structure coefficients. The structure coefficients revealed a more complex picture of the variables that contributed to the prediction of achievement. It was revealed that the FAT Math Posttest (r_s=.840) and the FAT Math Pretest (r_s=.781) were the strongest predictors followed by the FAT LA Posttest (r_s=.644) and the FAT LA Pretest (r_s=.589).

The difference between the predictive power of the FAT Math Pretest and Posttest appeared negligible.

Research Question 3: What are the statistically significant student variables that explain the largest amount of variance in student language arts achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Null Hypothesis 3: There are no statistically significant, research demonstrated, student variables that predict student language arts achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Answer: The null hypothesis is rejected. For School A the student variable gender, was a statistically significant predictor. The gender variable had a standardized beta of -.152, with reported statistical significance of p=.004 and a *t* value of 2.946. Therefore, it can be interpreted that being a male had a negative influence on language arts achievement as measured by the NJ ASK8.

The structure coefficient calculations supported the beta found for gender. Gender produced a structure coefficient of .389 (being a female improved performance). However, the student variable of attendance also emerged as a predictor based on the structure coefficients, -.333. Poor attendance influenced achievement negatively. SES might have acted as a suppressor variable in this case with a small structure coefficient of -.132. Howell (2002) described an example attributed to Jacob Cohen of a timed test in US history:

> We want to predict knowledge of historical facts. We give a test which supposedly tests that. But some people will do badly just because they read very slowly, and don't get through the exam. Others read very quickly, and do all the

> questions. We don't think that reading speed has anything to do with how much history you know, but it does affect your score. We want to "adjust" scores for reading speed, which is like saying 'The correlation between true historical knowledge and test score, *controlling for* reading speed.'(p.7)

SES has a long history of influencing student achievement. In a multiple regression model, SES also "drags" on other student achievement variables that could relate to achievement in the NJ ASK. In other words, SES suppresses the predictive power of other variables and thus, itself is an indirect predictor of achievement in this case.

Research Question 4: What are the statistically significant student variables that explain the largest amount of variance in student mathematics achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Null Hypothesis 4: There are no statistically significant, research demonstrated, student variables that predict student mathematics achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Answer: In terms of statistically significant betas, I fail to reject the null hypothesis. None of the student variables (gender, attendance, or SES) for School A were statistically significant in predicting student mathematics achievement by the state mandated NJ ASK8 for the 2008-2009 school year.

Although there were no statistically significant betas observed, the structure coefficients provided some potential insight into which student variables influenced the mathematics achievement. Attendance (r_s=.-342), SES (r_s= -.154), and gender (r_s=.128) all contributed to the NJ ASK8 Mathematics achievement. The results suggested that poor attendance had a negative drag on achievement, as does SES (which might have acted as a suppressor on other variables). It also appeared as if the female gender was a predictor of student achievement as measured by the NJ ASK8.

Research Question 5: What are the statistically significant school variables that explain the largest amount of variance in student language arts achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Null Hypothesis 5: There are no statistically significant, research demonstrated, school variables that predict student language arts achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Answer: The null hypothesis is rejected. For School A, student NJ ASK8 Math was a statistically significant predictor. As stated earlier, I believe NJ ASK8 Math is a spurious finding.

Research Question 6: What are the statistically significant school variables that explain largest amount of variance in student mathematics achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Null Hypothesis 6: There are no statistically significant, research demonstrated, school variables that predict student mathematics achievement as measured by the state

mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Answer: The null hypothesis is rejected. Student NJ ASK8 LA was a statistically significant predictor ($p<.001$) with a beta of .358 and a *t* value of 4.904.

The structure coefficient for the NJ ASK8 LA test supports the beta results as it was the second largest ($r_s=.809$) contributor to achievement.

School A Language Arts and Mathematics Summary

Language arts summary School A.

Statistically significant predictors for NJ ASK8 LA performance were: (a) NJ ASK8 Math, .363, $p<.001$, *t* value of 4.904 (b) FAT LA Posttest, .275, $p<.001$, *t* value of 4.358 (c) FAT LA Pretest, .248, $p<.001$, *t* value of 3.955, and (d) gender, .152, $p=.004$, *t* value of 2.946. The betas suggested that being female and performing well on the Pretest and Posttest FAT assessments will predict a large amount of the variance in NJ ASK8 LA performance.

The structure coefficients confirmed that the FAT LA Posttest ($r_s=.806$), FAT LA Pretest ($r_s=.771$), and gender ($r_s=.389$) were contributing to the prediction of achievement on the NJ ASK8 LA. Student attendance also emerged as a predictor based on the structure coefficients, -.333. Poor attendance influenced achievement negatively. Together, the betas and structure coefficients suggested that being female, with good attendance, and performing well on the FAT Pretest and Posttest assessments will predict the majority of the variance in NJ ASK8 LA performance.

The betas and structure coefficients suggested that the FAT LA Pretest results have approximately the same predictive power in terms of betas and structure coefficients as the FAT LA Posttest results.

Mathematics summary School A.

Two variables were statistically significant predictors variables of NJ ASK8 Mathematics performance: (a) FAT Math Posttest with a beta of .383, ($p<.001$) with a *t* value of 5.296, and (b) NJ ASK LA ($p<.001$) with a beta of .358 with a *t* value of 4.904. The structure coefficient for the NJ ASK8 LA test supports the beta results as it was the second largest (r_s=.809) contributor to achievement. The structure coefficients also provided a more complete picture of the variables that contributed to the prediction of achievement. It was revealed that FAT Math Posttest (r_s=.840) was the strongest predictor followed by NJ ASK8 LA (r_s=.809), then FAT Math Pretest (r_s=.781), the FAT LA Posttest (r_s=.644), and finally the FAT LA Pretest (r_s=.589). Intuitively, the structure coefficient results make sense. A student who performs well on those assessments will most likely perform well on the NJ ASK8 Math. Although the results are not surprising, they do provide some additional useful information. The FAT Math Posttest and Pretest are similar in their predictive qualities, (r_s=.840) versus (r_s=.781), as are the FAT LA Posttest and Pretest, (r_s=.644) versus (r_s=.589).

Common variables for School A.

Common variables emerged that affected both language arts and mathematics for School A. When taking into account both beta weights and structure coefficients the statistically significant variables were the respective FAT Pretest and Posttest for each subject area, language arts and math.

It is important to note when reviewing the results for School A that there is no certainty that the results are comparable to the other schools because School A did not have the ASI support that the other schools did. Therefore, there is no certainty that not having ASI in some way influenced the results for School A. The question remains, what would have had the largest influence on achievement if School A had an ASI program?

Language Arts and Mathematics Models for School B

I first created a matrix scatterplot to determine if the variables were related to each other in a linear fashion and found that to be true (see Figure 4). The data are arranged in columns because dichotomous variables are plotted according to data points. "Linearity would be violated if the data points bunch at the center of one column and at the ends of the other column" (Leech, Barrett, & Morgan, 2008, p. 103). The scatterplot results suggested that the assumption of linearity was not violated.

Language arts.

A multiple regression analysis for School B was performed between the dependent variable (NJ ASK8 LA scores) and the independent variables (gender, NJ ASK8 Math scores, attendance, SES, ASI, FAT LA and Math Pretest scores, and FAT LA and Math Posttest scores). I performed a descriptive analysis initially (see Table 16).

Using the enter method, a statistically significant model emerged (F = 36.329, p < .001, Adjusted R^2 .639.) The model summary suggested that approximately 64% of the variance in student performance on the NJ ASK8 LA and coefficients revealed four statistically significant predictors (see Tables 17 & 18).

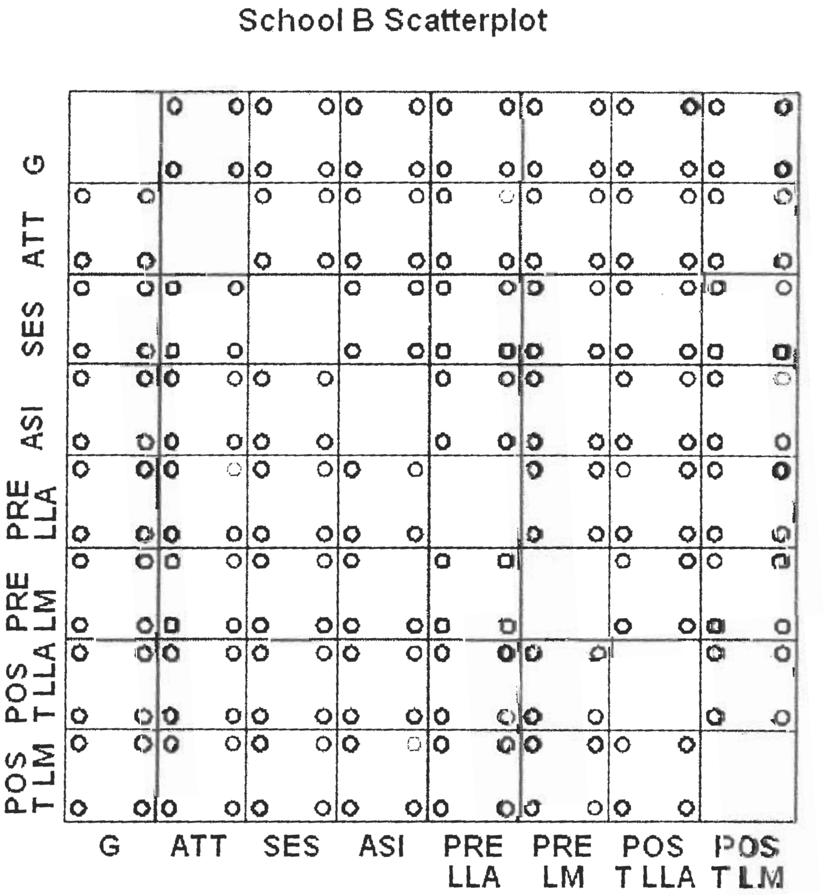


Figure 4. School B scatterplot.

Table 16

Descriptive Statistics for School B Language Arts

	Mean	Std. Deviation	N
ASK_LA	231.1878	19.12294	181
Gender 0=m	.5193	.50101	181
Attend	.0773	.26788	181
SES 0= Not	.1160	.32114	181
ASI 0=not	.1823	.38718	181
PRELA 0=Not Prof	.2762	.44838	181
PreM 0 = Not Prof	.3260	.47004	181
PostLA 0=Not Prf	.5691	.49658	181
PostM 0=Not Pr	.3867	.48835	181
ASK_M	242.7293	37.90234	181

Table 17

Model Summary School B Language Arts

Model					Change Statistics				
	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.810[a]	.657	.639	11.49728	.657	36.329	9	171	.000

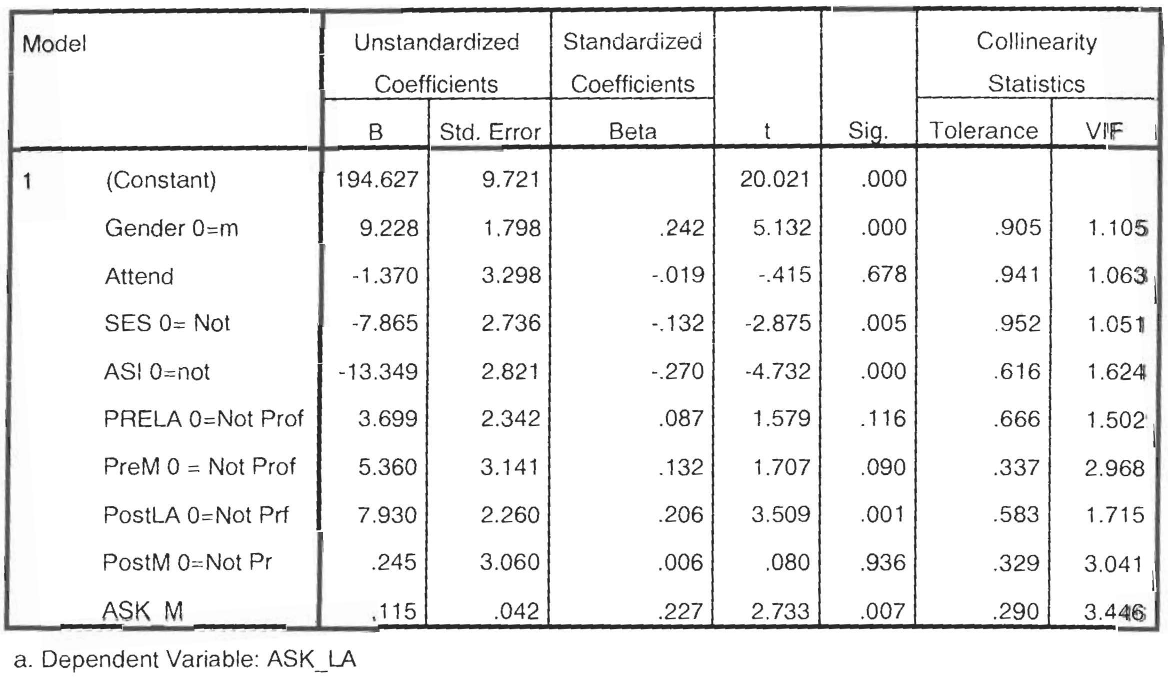
a. Predictors: (Constant), ASK_M, Gender 0=m, SES 0= Not, Attend , PRELA 0=Not Prof, ASI 0=not, PostLA 0=Not Prf, PreM 0 = Not Prof, PostM 0=Not Pr

b. Dependent Variable: ASK_LA

In this model, the variance explained (adj R^2) was .639, which indicated that approximately 64% of the variance in NJ ASK8 LA scores was explained by the variables in the model. Statistically significant predictors variables, their betas, p values, and *t* values were as follows: (a) ASI,- .270, p<.001, *t* value of 4.732, (b) gender, .242, p<.001, *t* value of 5.132, (c) NJ ASK8 Math, . 227, p<.001, *t* value 2.773, (d) FAT LA Posttest, .206, p=.001, *t* value of 3.509 , and (e) student SES, -.132, p=.005, *t* value of - 2.875. Multicollinearity was not an issue as the tolerance values for the predictors were not exceedingly low (<1-R^2) for any variable except NJ ASK8 Math. However, as stated earlier in the result for School A, the finding for the strong prediction value of NJ ASK8 Math for NJ ASK8 LA is spurious.

Table 18

Coefficients for School B Language Arts



Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	194.627	9.721		20.021	.000		
	Gender 0=m	9.228	1.798	.242	5.132	.000	.905	1.105
	Attend	-1.370	3.298	-.019	-.415	.678	.941	1.063
	SES 0= Not	-7.865	2.736	-.132	-2.875	.005	.952	1.051
	ASI 0=not	-13.349	2.821	-.270	-4.732	.000	.616	1.624
	PRELA 0=Not Prof	3.699	2.342	.087	1.579	.116	.666	1.502
	PreM 0 = Not Prof	5.360	3.141	.132	1.707	.090	.337	2.968
	PostLA 0=Not Prf	7.930	2.260	.206	3.509	.001	.583	1.715
	PostM 0=Not Pr	.245	3.060	.006	.080	.936	.329	3.041
	ASK_M	.115	.042	.227	2.733	.007	.290	3.446

a. Dependent Variable: ASK_LA

Mathematics.

A second multiple regression analysis for School B was performed between the dependent variable (NJ ASK8 Math scores) and the independent variables (gender, NJ ASK8 LA scores, attendance, SES, ASI, FAT LA and Math Pretest scores, and FAT LA and Math Posttest scores. I performed a descriptive analysis initially (see Table 19).

Using the enter method, a statistically significant model emerged (F = 49.333, p < .001, Adjusted R^2 .707.) The model summary suggested that approximately 71% of the variance in student performance on the NJ ASK8 Math and coefficients revealed four statistically significant predictors (see Tables 20 & 21).

Table 19

Descriptive Statistics for School B Mathematics

	Mean	Std. Deviation	N
ASK_M	242.7293	37.90234	181
Gender 0=m	.5193	.50101	181
Attend	.0773	.26788	181
SES 0= Not	.1160	.32114	181
ASI 0=not	.1823	.38718	181
PRELA 0=Not Prof	.2762	.44838	181
PreM 0 = Not Prof	.3260	.47004	181
PostLA 0=Not Prf	.5691	.49658	181
PostM 0=Not Pr	.3867	.48835	181
ASK_LA	231.1878	19.12294	181

Table 20

Model Summary School B Mathematics

Model					Change Statistics				
	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.850[a]	.722	.707	20.50528	.722	49.333	9	171	.000

a. Predictors: (Constant), ASK_LA, Attend , SES 0= Not, Gender 0=m, PRELA 0=Not Prof, ASI 0=not, PostM 0=Not Pr, PostLA 0=Not Prf, PreM 0 = Not Prof

b. Dependent Variable: ASK_M

In this model, the variance explained (adj R^2) was .707, which indicated that approximately 71% of the variance in NJ ASK8 Math scores was explained by the variables in the model. Statistically significant predictors variables, their betas, p values, and *t* values were as follows: (a) ASI,- .279, p<.001, *t* value of -5.539, (b) FAT Math Posttest, .240, p=001, *t* value of 3.541, (c) FAT Math Pretest, . 211, p=002, *t* value 3.093, (d) NJ ASK8 LA, .184, p=.007, *t* value of 2.733 , and (e) FAT LA Pretest,.115, p=.02, *t* value of 3.351. Multicollinearity was not an issue as the tolerance values for the predictors were not low (<1-R^2).

Table 21

Coefficients for School B Mathematics

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	146.801	29.651		4.951	.000		
	Gender 0=m	-5.459	3.420	-.072	-1.596	.112	.796	1.257
	Attend	-9.046	5.844	-.064	-1.548	.123	.953	1.049
	SES 0= Not	2.361	4.992	.020	.473	.637	.909	1.100
	ASI 0=Not	-27.291	4.927	-.279	-5.539	.000	.642	1.558
	PRELA 0=Not Prof	9.738	4.141	.115	2.351	.020	.677	1.476
	PreM 0 = Not Prof	17.005	5.498	.211	3.093	.002	.350	2.859
	PostLA 0=Not Prf	7.665	4.132	.100	1.855	.065	.555	1.802
	PostM 0=Not Pr	18.654	5.268	.240	3.541	.001	.353	2.834
	ASK_LA	.365	.134	.184	2.733	.007	.358	2.790

a. Dependent Variable: ASK_M

Research Questions and Answers for School B

Research Question 1: What is the strength and direction of the relationship between the proficiency categorizations of students in Grade 8 measured by commercially produced standardized formative assessment in language arts and the proficiency categorizations of students in Grade 8 measured by the language arts portion of the NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Null Hypothesis 1: No statistically significant relationship exists between a commercially produced standardized formative assessment tool proficiency categorization and students' language arts proficiency categorization on the NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Answer: The null hypothesis is rejected. In this model, the variance explained (adj R^2) was .639, which indicated that approximately 64% of the variance in NJ ASK8 LA scores by the variables in the model. The commercially prepared FAT LA Posttest was a statistically significant (p=.001) predictor variable with a beta of .206 and a *t* value of 3.509.

The structure coefficients support the FAT LA Posttest as the strongest predictor (r_s=.747) followed by the FAT LA Pretest (r_s=.563).

The results suggested that the FAT LA Posttest results appeared to be the best predictor of student NJ ASK8 LA achievement.

Research Question 2: What is the strength and direction of the relationship between the proficiency categorizations of students in Grade 8 measured by commercially produced standardized formative assessment in mathematics and the proficiency categorizations of students in Grade 8 measured by the mathematics portion of the NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Null Hypothesis 2: No statistically significant relationship exists between a commercially produced standardized formative assessment tool proficiency categorization and students' mathematics proficiency categorization on the NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Answer: The null hypothesis is rejected. The commercially prepared FAT Math Pretest was a statistically significant (p=.002) predictor variable with a beta of .211 and a *t* value of 3.093.

The structure coefficients revealed that both the FAT Math Posttest (r_s=.828) and the Pretest (r_s=.811) were strong predictors. FAT LA Posttest (r_s=.675) and Pretest (r_s=.586) were also predictors.

The difference between the predictive value of the FAT Math Posttest and Pretest appears to be negligible based on the structure coefficients.

Research Question 3: What are the statistically significant student variables that explain the largest amount of variance in student language arts achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Null Hypothesis 3: There are no statistically significant, research demonstrated, student variables that predict student language arts achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Answer: The null hypothesis is rejected. There were two statistically significant student variables that predicted achievement on the NJ ASK8 LA in School B: (a) gender, .242, (p<.001) with a *t* value of 5.132 and (b) student SES, -.132, (p=.005) with a *t* value of -2.875.

The structure coefficients revealed that gender was the strongest predictor (r_s=.392), followed by SES (r_s= -.314), and finally attendance (r_s= -.149). This can be interpreted that for School B females tended to outperform males on the NJ ASK8 LA assessment. Students who violated the attendance policy and that were eligible for free or reduced lunch performed poorer on the NJ ASK8 LA than their peers who did not violate the attendance policy and who were not eligible for free or reduced lunch.

Research Question 4: What are the statistically significant student variables that explain the largest amount of variance in student mathematics achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Null Hypothesis 4: There are no statistically significant, research demonstrated, student variables that predict student mathematics achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Answer: I fail to reject the null hypothesis. There were no statistically significant, research demonstrated, student variables that predicted student mathematics achievement. The structure coefficient for gender suggests that gender might be a suppressor variable in this case (r_s= - .039). Attendance (r_s= -.170) and SES (r_s= -.170) were negatively associated with achievement.

Research Question 5: What are the statistically significant school variables that explain the largest amount of variance in student language arts achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Null Hypothesis 5: There are no statistically significant, research demonstrated, school variables that predict student language arts achievement as measured by the state

mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Answer: The null hypothesis is rejected. One school variable was a statistically significant ($p<. 001$) predictor of achievement on the NJ ASK8 LA test: (a) ASI, -.270 and a *t* value of -4.732. The NJ ASK8 Math results were also statistically significant, but as stated earlier, I believe those findings to be spurious.

The structure coefficient of ASI was (r_s= -.713). The beta and structure coefficients suggest that being eligible for ASI services has a negative relationship to NJ ASK8 LA achievement.

Research Question 6: What are the statistically significant school variables that explain the largest amount of variance in student mathematics achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Null Hypothesis 6: There are no statistically significant, research demonstrated, school variables that predict student mathematics achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Answer: The null hypothesis is rejected. For School B the statistically significant school variables were (a) ASI,- .279, ($p<.001$) with a *t* value of -5.539 and (b) NJ ASK8 LA, .184, ($p=.007$) with a *t* value of 2.733.

The structure coefficients for ASI (r_s= -.695) and NJ ASK8 LA (r_s= .792) support the beta results. Being eligible for ASI services has a negative relationship to achievement whereas doing well on the NJ ASK8 LA has a positive relationship to performance on the NJ ASK8 Math assessment.

School B Language Arts and Mathematics Summary

Language arts summary School B.

In this model, the variance explained (adj R^2) was .639, which indicated that approximately 64% of the variance in NJ ASK8 LA scores was explained by the variables in the model. Statistically significant predictors variables, their betas, p values, and *t* values were as follows: (a) ASI,- .270, p<.001, *t* value of 4.732, (b) gender, .242, p<.001, *t* value of 5.132, (c) NJ ASK Math, . 227, p<.001, *t* value 2.773(spurious finding), (d) FAT LA Posttest, .206, p=.001, *t* value of 3.509 , and (e) student SES, -.132, p=.005, *t* value of -2.875.

The betas suggest that not being in need of ASI services, being female, doing well on the FAT LA Posttest, and not being eligible for free or reduced lunch accounts for most of the variance in performance on the NJ ASK8 LA assessment.

The structure coefficients support the FAT LA Posttest as the strongest predictor (r_s=.747) followed by the FAT LA Pretest (r_s=.563). The structure coefficients revealed that gender was the strongest predictor (r_s=.392), followed by SES (r_s= -.314), and finally attendance (r_s= -.149). Being female had a positive relationship to achievement, whereas being absent more than the allowable total and being eligible for free or reduced lunch had a negative relationship to achievement.

Mathematics summary School B.

In this model, the variance explained (adj R^2) was .707, which indicated that approximately 71% of the variance in NJ ASK8 Math scores was explained by the variables in the model. Statistically significant predictors variables, their betas, p values, and *t* values were as follows: (a) ASI,- .279, p<.001, *t* value of -5.539, (b) FAT Math Posttest, .240, p=001, *t* value of 3.541, (c) FAT Math Pretest, . 211, p=002, *t* value 3.093, (d) NJ ASK8 LA, .184, p=.007, *t* value of 2.733 , and (e) FAT LA Pretest,.115, p=.02, *t* value of 3.351.

The structure coefficients revealed that both the FAT Math Posttest (r_s=.828) and the Pretest (r_s=.811) were strong predictors, followed by NJ ASK8 LA (r_s= .792) and ASI (r_s= -.695). The FAT LA Posttest (r_s=.675) and Pretest (r_s=.586) were also strong predictors. The structure coefficient for gender suggested that gender might be a suppressor variable in this case (r_s= - .039) and attendance (r_s= -.170) and SES (r_s= -.170) were negatively associated with achievement. Performance on the FAT Math Posttest and Pretest, and not needing ASI services were the strongest predictors. The NJ ASK8 LA, FAT LA Posttest and Pretest were also strong predictors. Exceeding 16 absences and being eligible for free or reduced lunch were weaker predictors of NJ ASK8 Math achievement.

The predictive value of the FAT LA Pretest and Posttest were similar as were the FAT Math Pretest and Posttest.

Common variables for School B.

Common variables emerged that affected both language arts and mathematics for School B. When taking into account both beta weights and structure coefficients the statistically significant variables were the respective FAT Pretest and Posttest for each subject area, language arts and math as well as ASI, SES, and attendance.

Language Arts and Mathematics Models for School C

I first created a matrix scatterplot to determine if the variables were related to each other in a linear fashion and found that to be true (see Figure 5). The data are arranged in columns because dichotomous variables are plotted according to data points. "Linearity would be violated if the data points bunch at the center of one column and at the ends of the other column" (Leech, Barrett, & Morgan. 2008, p. 103). The scatterplot results suggested that the assumption of linearity was not violated.

Language arts.

A multiple regression analysis for School C was performed between the dependent variable (NJ ASK8 LA scores) and the independent variables (gender, NJ ASK8 Math scores, attendance, SES, ASI, FAT LA and Math Pretest scores, and FAT LA and Math Posttest scores). I performed a descriptive analysis initially (see Table 22).

Using the enter method, a statistically significant model emerged (F = 20.246, p < .001, Adjusted R^2.614.) The model summary suggested that approximately 61% of the variance in student performance on the NJ ASK8 LA and coefficients revealed four statistically significant predictors (see Tables 23 & 24).

In this model, the variance explained (adj R^2) was .614, which indicated that approximately 61% of the variance in NJ ASK8 LA scores was explained by the variables in the model. Statistically significant predictor variables, their betas, p values, and t values were as follows: (a) NJ ASK8 Math .703, p<.001, t value of 7.310, (b) FAT LA Pretest, .214, p=.003, t value of 3.067, (c) FAT LA Posttest, .212, p=.005, t value of 2.884, and FAT Math Posttest, -.189, p=.021, t value -2.347.

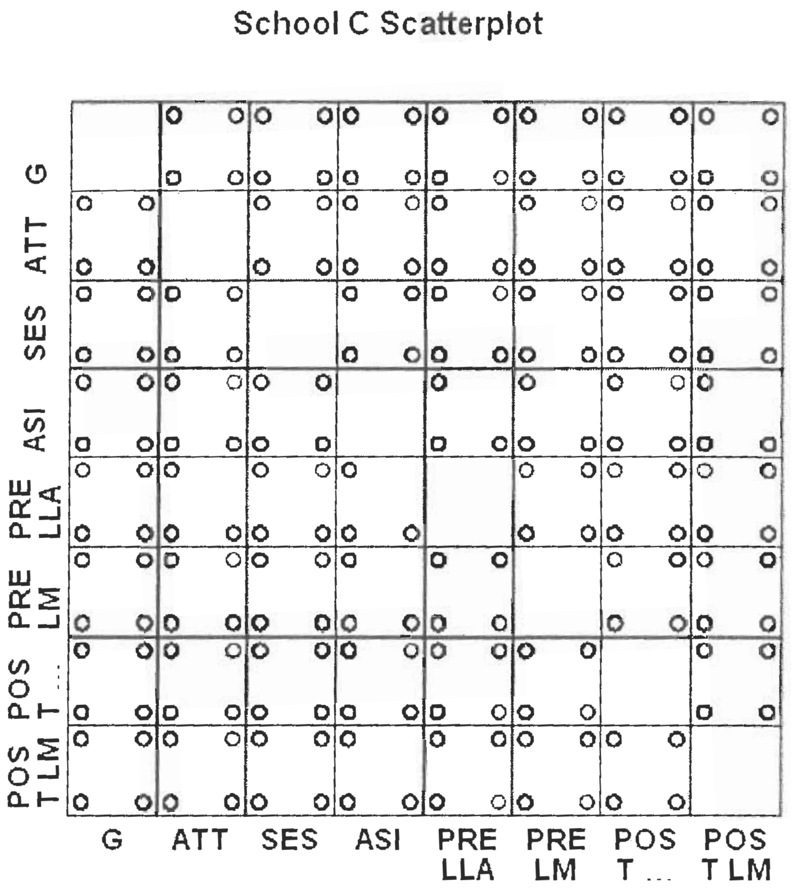


Figure 5. School C scatterplot.

Table 22

Descriptive Statistics for School C Language Arts

	Mean	Std. Deviation	N
ASK_LA	224.5000	19.38256	110
Gender 0=m	.5091	.50221	110
Attend	.1273	.33480	110
SES 0= Not	.4273	.49695	110
ASI 0=not	.1545	.36313	110
PRELA 0=Not Prof	.1091	.31318	110
PreM 0 = Not Prof	.2182	.41490	110
PostLA 0=Not Prf	.3545	.48056	110
PostM 0=Not Pr	.3545	.48056	110
ASK_M	237.8182	35.96308	110

Table 23

Model Summary for School C Language Arts

Model					Change Statistics				
	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.804[a]	.646	.614	12.04578	.646	20.246	9	100	.000

a. Predictors: (Constant), ASK_M, Attend , SES 0= Not, Gender 0=m, PRELA 0=Not Prof, ASI 0=not, PostLA 0=Not Prf, PostM 0=Not Pr, PreM 0 = Not Prof

b. Dependent Variable: ASK_LA

Multicollinearity was not an issue as the tolerance values for the predictors were not exceedingly low (<1-R^2) for every variable except NJ ASK8 Math with a reported value of .383, which was marginal.

However, as stated earlier in the result for School A, the finding for the strong prediction value of NJ ASK8 Math for NJ ASK8 LA is spurious but because of the strong results for the NJ ASK8 Math and the FAT Math Pretest I ran the model without NJ ASK8 Math.

Table 24

Coefficients for School C Language Arts Model 1

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	127.897	12.255		10.436	.000		
	Gender 0=m	4.717	2.537	.122	1.859	.066	.820	1.220
	Attend	5.289	3.520	.091	1.502	.136	.958	1.043
	SES 0= Not	2.065	2.453	.053	.842	.402	.896	1.116
	ASI 0=Not	1.893	3.696	.035	.512	.610	.739	1.353
	PRELA 0=Not Prof	13.249	4.319	.214	3.067	.003	.727	1.375
	PreM 0 = Not Prof	1.977	3.960	.042	.499	.619	.493	2.028
	PostLA 0=Not Prf	8.533	2.958	.212	2.884	.005	.659	1.518
	PostM 0=Not Pr	-7.637	3.253	-.189	-2.347	.021	.545	1.836
	ASK_M	.379	.052	.703	7.310	.000	.383	2.613

a. Dependent Variable: ASK_LA

Table 25

Corrected Model for School C Language Arts

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	215.689	3.010		71.669	.000		
	Gender 0=m	-1.427	2.951	-.037	-.484	.630	.921	1.086
	Attend	6.274	4.336	.108	1.447	.151	.960	1.042
	SES 0= Not	1.576	3.022	.040	.522	.603	.897	1.115
	ASI 0=Not	-7.395	4.278	-.139	-1.729	.087	.838	1.193
	PRELA 0=Not Prof	16.591	5.294	.268	3.134	.002	.736	1.359
	PreM 0 = Not Prof	9.057	4.732	.194	1.914	.058	.525	1.906
	PostLA 0=Not Prf	13.130	3.563	.326	3.685	.000	.690	1.450
	PostM 0=Not Pr	2.166	3.654	.054	.593	.555	.656	1.524

a. Dependent Variable: ASK_LA

I reran the model to check the influence of the variable and discovered the adjusted R^2 was .414 with an F statistic of 10.595.

Mathematics.

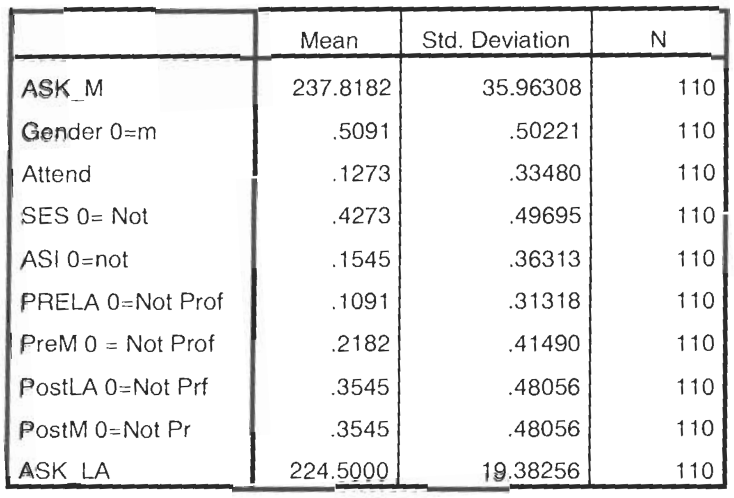
A second multiple regression analysis for School C was performed between the dependent variable (NJ ASK8 Math scores) and the independent variables (gender, NJ ASK8 LA scores, attendance, SES, ASI, FAT LA and Math Pretest scores, and FAT LA and Math Posttest scores. I performed a descriptive analysis initially (see Table 26).

Using the enter method, a statistically significant model emerged (F = 33.437, $p < .001$, Adjusted R^2 .728.) The model summary suggested that approximately 73% of the variance in student performance on the NJ ASK8 Math and coefficients revealed four statistically significant predictors (see Tables 27 & 28).

In this model, the variance explained (adj R^2) was .728, which indicated that approximately 73% of the variance in NJ ASK8 Math scores was explained by the variables in the model.

Table 26

Descriptive Statistics for School C Mathematics



	Mean	Std. Deviation	N
ASK_M	237.8182	35.96308	110
Gender 0=m	.5091	.50221	110
Attend	.1273	.33480	110
SES 0= Not	.4273	.49695	110
ASI 0=not	.1545	.36313	110
PRELA 0=Not Prof	.1091	.31318	110
PreM 0 = Not Prof	.2182	.41490	110
PostLA 0=Not Prf	.3545	.48056	110
PostM 0=Not Pr	.3545	.48056	110
ASK_LA	224.5000	19.38256	110

Table 27

Model Summary for School C Mathematics

Model					Change Statistics				
	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.866[a]	.751	.725	18.75148	.751	33.437	9	100	.000

a. Predictors: (Constant), ASK_LA, Gender 0=m, Attend, SES 0=Not, ASI 0=Not, Post M 0=Not Pr, PRELA 0=Not Prof, Post LA 0=Not, PreM 0=Not Prof

b. Dependent Variable: ASK_M

Statistically significant predictor variables, their betas, p values, and *t* values were as follows: (a) NJ ASK8 LA, .495, p<.001, *t* value of 7.310, (b) FAT Math Posttest, .319, p<.001, *t* value of 5.163, (c) gender, -.208, p<. 001, *t* value -3.992, and (d) ASI, -.179, *t* value -3.230. Multicollinearity was not an issue as the tolerance values for the predictors were not low ($<1-R^2$).

Table 28

Coefficients for School C Mathematics

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	33.424	27.371		1.221	.225		
	Gender 0=m	-14.896	3.731	-.208	-3.992	.000	.919	1.089
	Attend	-3.166	5.532	-.029	-.572	.568	.940	1.064
	SES 0= Not	-2.738	3.822	-.038	-.716	.475	.894	1.118
	ASI 0=Not	-17.707	5.482	-.179	-3.230	.002	.814	1.229
	PRELA 0=Not Prof	-6.426	7.004	-.056	-.917	.361	.670	1.491
	PreM 0 = Not Prof	10.354	6.084	.119	1.702	.092	.506	1.975
	PostLA 0=Not Prf	.063	4.793	.001	.013	.989	.608	1.645
	PostM 0=Not Pr	23.867	4.622	.319	5.163	.000	.654	1.530
	ASK_LA	.919	.126	.495	7.310	.000	.544	1.839

a. Dependent Variable: ASK_M

Research Questions and Answers for School C

Research Question 1: What is the strength and direction of the relationship between the proficiency categorizations of students in Grade 8 measured by commercially produced standardized formative assessment in language arts and the proficiency categorizations of students in Grade 8 measured by the language arts portion of the NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Null Hypothesis 1: No statistically significant relationship exists between a commercially produced standardized formative assessment tool proficiency categorization and students' language arts proficiency categorization on the NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Answer: The null hypothesis is rejected. In this model, the variance explained (adj R^2) was .614, which indicated that approximately 64% of the variance in NJ ASK8 LA scores was explained by the variables in the model. The following variables had statistically significant betas: (a) FAT LA Posttest, .326, p<.001, *t* value of 3.685, and (b) FAT LA Pretest, .268, p=.002, *t* value of 3.134.

The structure coefficients support the FAT LA Posttest (r_s= -.658) and Pretest (r_s= -.574) were the strongest predictors. The Posttest coefficient suggested that it was a better predictor than the pretest.

Research Question 2: What is the strength and direction of the relationship between the proficiency categorizations of students in Grade 8 measured by commercially produced standardized formative assessment in mathematics and the proficiency categorizations of students in Grade 8 measured by the mathematics portion of the NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Null Hypothesis 2: No statistically significant relationship exists between a commercially produced standardized formative assessment tool proficiency categorization and students' mathematics proficiency categorization on the NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Answer: The null hypothesis is rejected. The commercially prepared FAT Math Posttest was a statistically significant ($p<.001$) predictor variable with a beta of .319 and a *t* value of 5.163.

The structure coefficients revealed a more complex picture of which variables were contributing to the prediction of achievement. By computing and analyzing the structure coefficients it was revealed that the FAT LA Posttest (r_s= -.720) and Pretest (r_s= -.657) appeared to be strong predictors. The results suggested that administering both tests as a predictive instrument might not be necessary.

Research Question #3: What are the statistically significant student variables that explain the largest amount of variance in student language arts achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Null Hypothesis 3: There are no statistically significant, research demonstrated, student variables that predict student language arts achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Answer: I fail to reject the null hypothesis. There were no statistically significant variables. However, a review of the structure coefficients suggested that low SES (-.204) contributed negatively to achievement on the NJ ASK8 LA. Gender (.014) and attendance (.023) appeared to be suppressor variables that might have masked some of the predictive power of the other variables.

Research Question 4: What are the statistically significant student variables that explain the largest amount of variance in student mathematics achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Null Hypothesis 4: There are no statistically significant, research demonstrated, student variables that predict student mathematics achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Answer: The null hypothesis is rejected. One student variable, gender, was statistically significant (p<. 001) with a beta of -.208 and a *t* value of -3.992. This can be interpreted that for School C females tended to outperform males on the NJ ASK8 Math assessment.

The structure coefficients confirmed gender (r_s= -.286) as a student variable that predicted achievement. However, low student SES was also almost as strong (r_s= -.243). It was revealed that males with a lower SES perform poorer on the NJ ASK8 Math test than students with a higher SES.

Research Question 5: What are the statistically significant school variables that explain the largest amount of variance in student language arts achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Null Hypothesis 5: There are no statistically significant, research demonstrated, school variables that predict student language arts achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Answer: I fail to reject the null hypothesis. There were no statistically significant predictors. However, the structure coefficients revealed that ASI was a strong predictor (r_s= -.384). Students that were eligible for ASI performed poorer on the NJ ASK8 LA than students that were not eligible for ASI.

Research Question 6: What are the statistically significant school variables that explain the largest amount of variance in student mathematics achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Null Hypothesis 6: There are no statistically significant, research demonstrated, school variables that predict student mathematics achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Answer: The null hypothesis is rejected. For School C the statistically significant school variable was (a) ASI,- .179 (p=.002) with a *t* value of -3.230.

School C Language Arts and Mathematics Summary

Language arts summary School C.

Statistically significant predictor variables, their betas, p values, and *t* values for the second model were as follows: (a) FAT LA Posttest, .326, p<.001, *t* value of 3.685, and (b) FAT LA Pretest, .268, p=.002, *t* value of 3.134. The betas suggested that the performance on the FAT Posttest and Pretest predicted the variance in performance on the NJ ASK8 LA test. These variables accounted for similar amounts of variance.

Once again, a review of the structure coefficients provided a more complete explanation of the variance. The structure coefficients supported the FAT LA Posttest (r_s= -.658) and Pretest (r_s= -.574) as the strongest predictors but also revealed that being eligible for ASI was a moderate negative predictor (r_s= -.384). Low SES (-.204) was also found to negatively predict achievement on the NJ ASK8 LA. Gender (.014) and attendance (.023) appeared to be suppressor variables that might have masked some of the predictive power of the other variables such as SES. For example, if low SES was highly correlated with attendance, attendance acting as a suppressor could have masked the influence of SES on the variance.

Mathematics summary School C.

Statistically significant predictor variables, their betas, p values, and *t* values were as follows: (a) NJ ASK8 LA, .495, p<.001, *t* value of 7.310, (b) FAT Math Posttest, .319, p<.001, *t* value of 5.163, (c) gender, -.208, p<. 001, *t* value -3.992, and (d) ASI, -.179, *t* value -3.230. The betas suggested that being female, not eligible for ASI services and doing well on the NJ ASK8 LA and FAT Math Posttest predicted positive performance on the NJ ASK8 Math test.

By computing and analyzing the structure coefficients it was revealed that the FAT LA Posttest (r_s= -.720) and Pretest (r_s= -.657) appeared to be strong predictors of student achievement. The structure coefficients confirmed gender (r_s= -.286) as a student variable that predicted achievement. However, low student SES was also almost as strong (r_s= -.243). Therefore, males with low SES tended to perform poorer on the NJ ASK 8 Math test.

Common variables for School C.

Common variables emerged that affected both language arts and mathematics for School C. When taking into account both beta weights and structure coefficients the statistically significant variables were the respective FAT Posttest for each subject area, language arts and math as well as ASI and SES.

Language Arts and Mathematics Model for School D

I first created a matrix scatterplot to determine if the variables were related to each other in a linear fashion and found that to be true (see Figure 6). The data are arranged in columns because dichotomous variables are plotted according to data points. "Linearity would be violated if the data points bunch at the center of one column and at the ends of the other column" (Leech, Barrett, & Morgan, 2008, p. 103). The scatterplot results suggested that the assumption of linearity was not violated.

Language arts.

A multiple regression analysis for School D was performed between the dependent variable (NJ ASK8 LA scores) and the independent variables (gender, NJ ASK8 Math scores, attendance, SES, ASI, FAT LA and Math Pretest scores, and FAT LA and Math Posttest scores). Analysis was performed using what SPSS refers to as

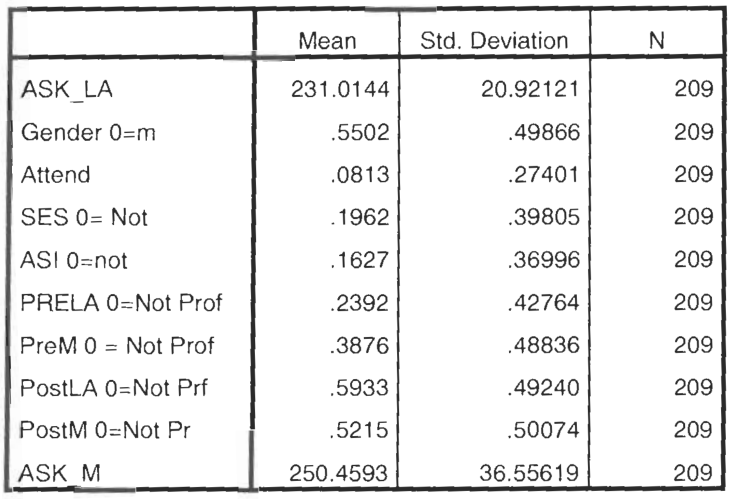
Enter (also known as simultaneous regression). I performed a descriptive analysis initially (see Table 29).

Using the enter method, a statistically significant model emerged (*F* = 30.316, p < .001, Adjusted R^2 .559.) The model summary suggested that approximately 56% of the variance in student performance on the NJ ASK8 LA and coefficients revealed four statistically significant predictors (see Tables 30 & 31).

In this model, the variance explained (adj R^2) was .559, which indicated that approximately 56% of the variance in NJ ASK8 LA scores was explained by the variables in the model.

Table 29

Descriptive Statistics for School D Language Arts



	Mean	Std. Deviation	N
ASK_LA	231.0144	20.92121	209
Gender 0=m	.5502	.49866	209
Attend	.0813	.27401	209
SES 0= Not	.1962	.39805	209
ASI 0=not	.1627	.36996	209
PRELA 0=Not Prof	.2392	.42764	209
PreM 0 = Not Prof	.3876	.48836	209
PostLA 0=Not Prf	.5933	.49240	209
PostM 0=Not Pr	.5215	.50074	209
ASK_M	250.4593	36.55619	209

Statistically significant predictor variables, their betas, p values, and *t* values were as follows: (a) NJ ASK8 Math .443, p<.001, *t* value of 5.956, (b) FAT LA Pretest, .187, p<.001, *t* value of 3.750, (c) ASI, -.119, p=.033, t value -2.141, (d) FAT LA Posttest, .113, p=.044, *t* value of 2.023, and (e) gender, .102, p=.033, t value -2.147.

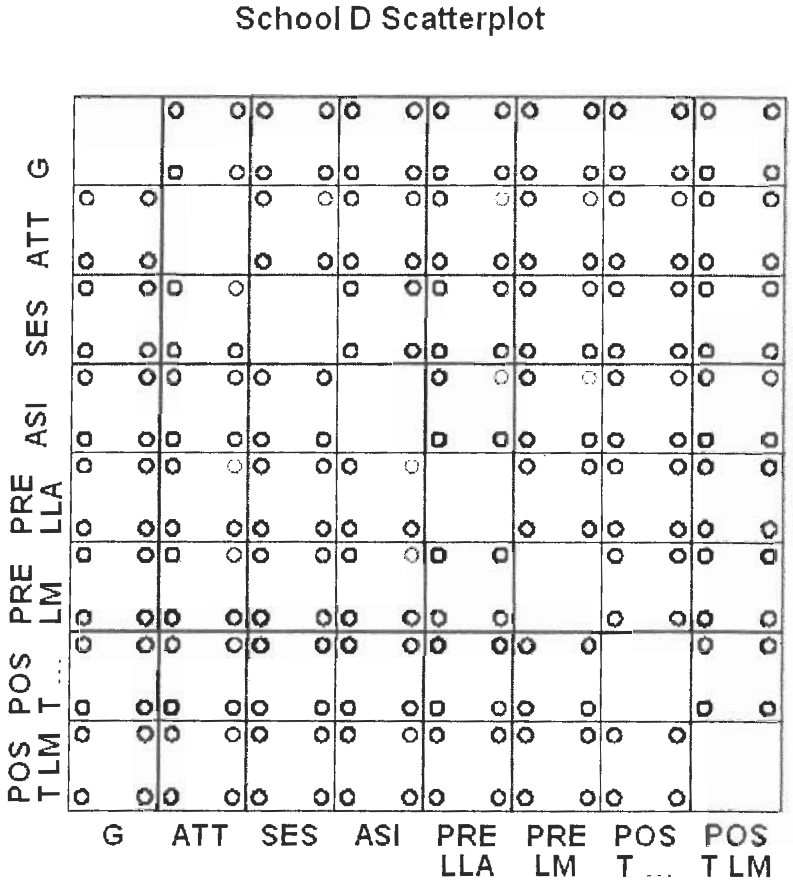


Figure 6. School D scatterplot.

Table 30

Model Summary for School D Language Arts

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.760[a]	.578	.559	13.89048	.578	30.316	9	199	.000

a. Predictors: (Constant), ASK_M, Gender 0=m, SES 0= Not, Attend , PRELA 0=Not Prof, PostLA 0=Not Prf, ASI 0=not, PostM 0=Not Pr, PreM 0 = Not Prof

b. Dependent Variable: ASK_LA

Table 31

Coefficients for School D Language Arts

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	158.059	10.093		15.661	.000		
	Gender 0=m	4.278	1.992	.102	2.147	.033	.940	1.064
	Attend	4.157	3.699	.054	1.124	.262	.903	1.107
	SES 0= Not	.781	2.579	.015	.303	.762	.880	1.136
	ASI 0=Not	-6.730	3.143	-.119	-2.141	.033	.686	1.458
	PRELA 0=Not Prof	9.142	2.438	.187	3.750	.000	.853	1.172
	PreM 0 = Not Prof	5.205	2.949	.122	1.765	.079	.447	2.235
	PostLA 0=Not Prf	4.783	2.365	.113	2.023	.044	.684	1.462
	PostM 0=Not Pr	1.259	2.602	.030	.484	.629	.546	1.830
	ASK_M	.254	.043	.443	5.956	.000	.383	2.611

a. Dependent Variable: ASK_LA

Mathematics.

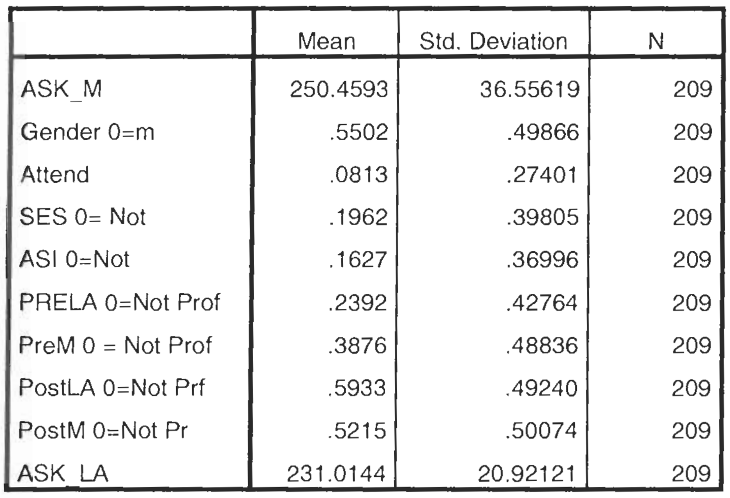
A second multiple regression analysis for School D was performed between the dependent variable (NJ ASK8 Math scores) and the independent variables (gender, NJ

ASK8 LA scores, attendance, SES, ASI, FAT LA and Math Pretest scores, and FAT LA and Math Posttest scores. I performed a descriptive analysis initially (see Table 32).

Using the enter method, a statistically significant model emerged (*F* = 45.910, p < .001, Adjusted R^2 .660.) The model summary suggested that approximately 66% of the variance in student performance on the NJ ASK8 Math and coefficients revealed four statistically significant predictors (see Tables 33& 34).

Table 32

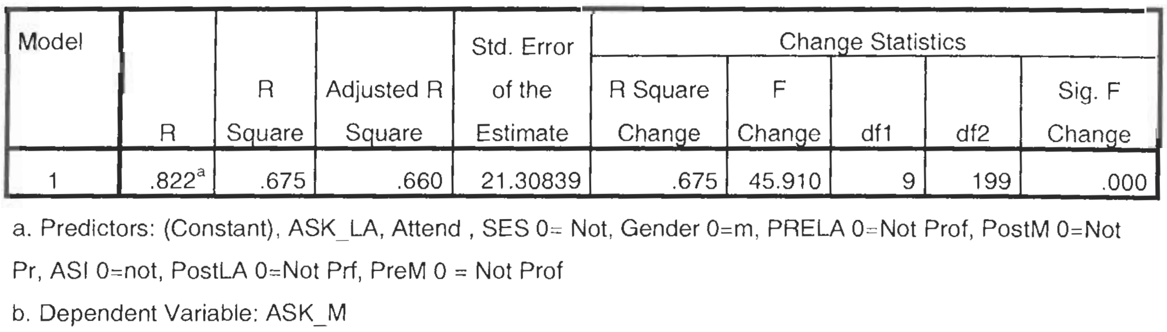
Descriptive Statistics for School D Mathematics



	Mean	Std. Deviation	N
ASK_M	250.4593	36.55619	209
Gender 0=m	.5502	.49866	209
Attend	.0813	.27401	209
SES 0= Not	.1962	.39805	209
ASI 0=Not	.1627	.36996	209
PRELA 0=Not Prof	.2392	.42764	209
PreM 0 = Not Prof	.3876	.48836	209
PostLA 0=Not Prf	.5933	.49240	209
PostM 0=Not Pr	.5215	.50074	209
ASK_LA	231.0144	20.92121	209

Table 33

Model Summary for School D Mathematics



Model				Std. Error	Change Statistics				
	R	R Square	Adjusted R Square	of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.822[a]	.675	.660	21.30839	.675	45.910	9	199	.000

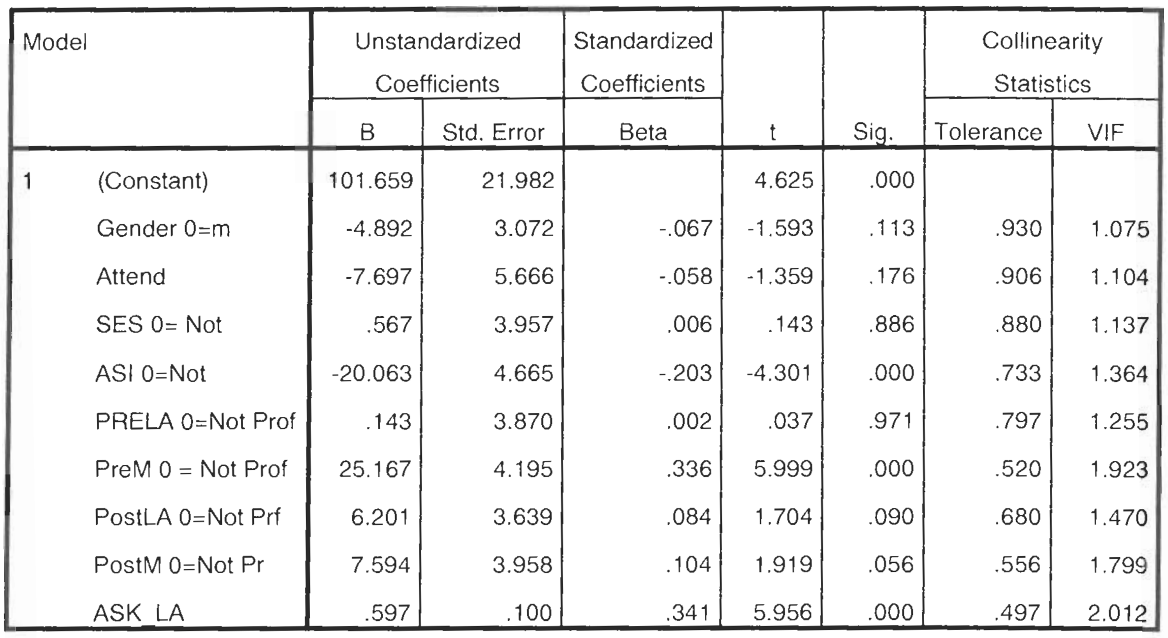
a. Predictors: (Constant), ASK_LA, Attend , SES 0= Not, Gender 0=m, PRELA 0=Not Prof, PostM 0=Not Pr, ASI 0=not, PostLA 0=Not Prf, PreM 0 = Not Prof

b. Dependent Variable: ASK_M

In this model, the variance explained (adj R^2) was .660, which indicated that approximately 66% of the variance in NJ ASK8 Math scores was explained by the variables in the model. Statistically significant predictor variables, their betas, p values, and *t* values were as follows: (a) NJ ASK8 LA, .341, p<.001, *t* value of 5.956, (b) FAT Math Pretest, .336, p<.001, *t* value of 5.999, (c) ASI, -.203, *t* value -4.301, (d) FAT Math Posttest, .104, p=.056, *t* value 1.919. Multicollinearity was not an issue as the tolerance values for the predictors were not low ($<1-R^2$).

Table 34

Coefficients for School D Mathematics



Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	101.659	21.982		4.625	.000		
	Gender 0=m	-4.892	3.072	-.067	-1.593	.113	.930	1.075
	Attend	-7.697	5.666	-.058	-1.359	.176	.906	1.104
	SES 0= Not	.567	3.957	.006	.143	.886	.880	1.137
	ASI 0=Not	-20.063	4.665	-.203	-4.301	.000	.733	1.364
	PRELA 0=Not Prof	.143	3.870	.002	.037	.971	.797	1.255
	PreM 0 = Not Prof	25.167	4.195	.336	5.999	.000	.520	1.923
	PostLA 0=Not Prf	6.201	3.639	.084	1.704	.090	.680	1.470
	PostM 0=Not Pr	7.594	3.958	.104	1.919	.056	.556	1.799
	ASK_LA	.597	.100	.341	5.956	.000	.497	2.012

a. Dependent Variable: ASK_M

Research Questions and Answers for School D

Research Question 1: What is the strength and direction of the relationship between the proficiency categorizations of students in Grade 8 measured by commercially produced standardized formative assessment in language arts and the scale scores of students in Grade 8 measured by the language arts portion of the NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Null Hypothesis 1: No statistically significant relationship exists between a commercially produced standardized formative assessment tool proficiency categorization and students' language arts proficiency categorization on the NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Answer: The null hypothesis is rejected. There were two statistically significant predictors: (a) FAT LA Pretest, .187, p<.001, *t* value of 3.750, and (b) FAT LA Posttest, .113, p=.044, *t* value of 2.023. The structure coefficients suggested that the FAT LA Posttest (r_s = .559) and FAT LA Pretest (r_s = .559) accounted for the same amount of variance.

Research Question 2: What is the strength and direction of the relationship between the proficiency categorizations of students in Grade 8 measured by commercially produced standardized formative assessment in mathematics and the proficiency categorizations of students in Grade 8 measured by the mathematics portion of the NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Null Hypothesis 2: No statistically significant relationship exists between a commercially produced standardized formative assessment tool proficiency categorization and students' mathematics proficiency categorization on the NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Answer: The null hypothesis is rejected. There were two statistically significant variables: (a) FAT Math Pretest, .336, p<.001, *t* value of 5.999, and (b) FAT Math

Posttest, .104, p= .056, *t* value of 1.919. The structure coefficients suggested that the FAT Math Pretest is the best predictor (r_s = .819) followed by FAT Math Posttest (r_s = .678).

Research Question 3: What are the statistically significant student variables that explain the largest amount of variance in student language arts achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Null Hypothesis 3: There are no statistically significant, research demonstrated, student variables that predict student language arts achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Answer: The null hypothesis is rejected. One student variable was a statistically significant predictor of achievement: gender had a standardized beta of .102, with an observed *t* value of 2.147, and a reported significance of .033.

The structure coefficients provided a more complete explanation. Student SES (r_s = -.212) was the strongest predictor, followed by gender (r_s =.198) and then attendance (r_s = -.141).

Research Question 4: What are the statistically significant student variables that explain the largest amount of variance in student mathematics achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Null Hypothesis 4: There are no statistically significant, research demonstrated, student variables that predict student mathematics achievement as measured by the state

mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Answer: I fail to reject the null hypothesis. There were no statistically significant student variables.

The structure coefficients suggested that attendance (r_s = -.249) and SES (r_s = -.221) were predictors. Students with a low SES who violated the attendance policy performed poorer on the NJ ASK8 Math test.

Research Question 5: What are the statistically significant school variables that explain the largest amount of variance in student language arts achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Null Hypothesis 5: There are no statistically significant, research demonstrated, school variables that predict student language arts achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Answer: The null hypothesis is rejected. For School D, ASI was a statistically significant (p= .033) predictor with a beta of -.119 and a *t* value of -2.141. The structure coefficient confirmed ASI as a strong predictor (r_s = -.614).

Research Question 6: What are the statistically significant school variables that explain the largest amount of variance in student mathematics achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district?

Null Hypothesis 6: There are no statistically significant, research demonstrated, school variables that predict student mathematics achievement as measured by the state mandated NJ ASK8 for the 2008-2009 school year for four of the middle schools in the district.

Answer: The null hypothesis is rejected. For School D, ASI was a statistically significant (p<.001) predictor with a beta of -.203 and a *t* value of -4.301. The structure coefficient suggested that ASI is a strong school predictor (r_s = -.642).

School D Language Arts and Mathematics Summary

Language arts summary School D.

In this model, the variance explained (adj R^2) was .559, which indicated that approximately 56% of the variance in NJ ASK8 LA scores was explained by the variables in the model. Statistically significant predictor variables, their betas, p values, and *t* values were as follows: (a) NJ ASK8 Math, .443, p<.001, *t* value of 5.956, (b) FAT LA Pretest, .187, p<.001, *t* value of 3.750, (c) ASI, -.119, p=.033, t value -2.141, (d) FAT LA Posttest, .113, p=.044, *t* value of 2.023, and (e) gender,.102, p=.033, t value -2.147.

The betas suggested that not being in need of ASI services, being female, and doing well on the FAT LA and Math Pretest and Posttests accounted for the most variance in performance on the NJ ASK8 LA test.

The structure coefficients supported the FAT LA Pretest as the strongest predictors (r_s = .559) along with the FAT LA Posttest (r_s = .559). The structure coefficients of gender was .198.

Mathematics summary School D.

In this model, the variance explained (adj R^2) was .660, which indicated that approximately 66% of the variance in NJ ASK8 Math scores was explained by the variables in the model. Statistically significant predictor variables, their betas, p values, and *t* values were as follows: (a) NJ ASK8 LA, .341, p<.001, *t* value of 5.956, (b) FAT Math Pretest, .336, p<.001, *t* value of 5.999, (c) ASI, -.203, *t* value -4.301, (d) FAT Math Posttest, .104, p=.056, *t* value 1.919.

The structure coefficients revealed that both FAT Math Pretest (r_s = .819) and the Posttest (r_s = .678) were the strongest predictors, followed by ASI (r_s = -.642). Performance on the FAT Math Pretest and Posttest, and not needing ASI services were the strongest predictors of student achievement on the NJ ASK8 Math test.

Common variables for School D.

Common variables emerged that affected both language arts and mathematics for School D. When taking into account both beta weights and structure coefficients the statistically significant variables were the respective FAT Posttest for each subject area, language arts and math as well as ASI.

Table 35 provides a breakdown by school of variables that were significant when analyzing betas and structure coefficients in predicting student achievement on the NJ ASK8.

Table 35

Significant Variables Breakdown by School

	School A		School B		School C		School D	
Variable	LA	Math	LA	Math	LA	Math	LA	Math
FAT Pre LA	X		X		X		X	
FAT Post LA	X		X		X		X	
FAT Pre Math		X		X				X
FAT Post Math		X		X		X		X
ASI	N/A	N/A	X	X	X	X	X	X
SES		X	X	X	X	X	X	X
Gender	X		X			X	X	
Attendance	X	X	X	X				X

Chapter V

CONCLUSIONS AND RECOMMENDATIONS

Introduction

The purpose of this study was to determine the strength and the direction of the relationships between student and school variables found in the extant literature to influence student achievement and aggregate district student NJ ASK scores in Grades 8 language arts and mathematics. By focusing on multiple school and student variables that significantly influence student achievement, I aimed to produce research-based evidence to assist all stakeholders in public education regarding the reform initiatives addressed herein. This study was guided by the following overarching research question: What student and school variables, found in the extant literature explain the greatest variance in student achievement on the NJ ASK8 language arts and mathematics sections?

The results of the study revealed each school produced a combination of site specific results and results common across sites regarding the strength of each independent variable to predict student achievement. Therefore, first I will discuss the conclusions for each variable and then present conclusions that relate to all sites in the study. I will present recommendations for policy and practice using the same format.

Formative/Interim Assessment Variable

Conclusions

As the New Jersey Department of Education continues to mandate requirements that it deems will help improve student achievement, these mandates warrant further discussion and investigation by stakeholders. The use of commercial products marketed as formative assessment tools such as the FAT program is encouraged by officials in the New Jersey Department of Education. In a memo released in 2008, former Commissioner of Education Lucille Davy wrote:

> Formative assessment resources allow educators to evaluate and measure student achievement continually, in a low-pressure context, using non-secure benchmark testing forms, item pools, distractor analysis, item authoring software, and associated score reports. Formative assessment resources allow teachers to connect specific grade level indicators with specific students or groups of students. They are formative in helping teachers shape and improve instruction as teachers shape and improve student understanding. (p.1)

According to The Company FAT is a "Formative Assessment System designed to improve instructional effectiveness in the classroom" (2009, p.1). To recall, for the purpose of this study the definition of "formative assessment" formulated by Perie, Marion, and Gong (2007) was used:

> Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes. Thus, it is done by the teacher in the classroom for the explicit purpose of diagnosing where students are in their learning, where gaps in knowledge and understanding exist, and how to help teachers and students improve student learning. The assessment is embedded within the learning activity and linked directly to the current unit of instruction…There is little interest or sense in trying to aggregate formative assessment information beyond the specific classroom. (p.1)

FAT provides the language arts and mathematics pretests and posttests, but not strategies or activities for enhancing or "improving instruction" as the Commissioner mentioned formative assessments should do. It appears that FAT, the manner in which it was commissioned in the district studied, aligns more with what Perie, Marion, and Gong (2007) would refer to as interim assessments rather than formative assessments:

> The assessments that fall between formative and summative assessments including the medium-scale, medium-cycle assessments currently in wide use. Interim assessments (1) evaluate students' knowledge and skills relative to a specific set of academic goals, typically within a limited time frame, and (2) are designed to inform decisions at both the classroom and beyond the classroom level, such as the school or district level. Thus, they may be given at the classroom level to provide information for the teacher, but unlike true formative assessments, they results of interim assessments can be meaningfully aggregated and reported at a broader level. As such, the timing of the administration is likely to be controlled by the school or district rather than by the teacher, which therefore makes these assessments less instructionally relevant than formative assessments. These assessments may serve a variety of purposes, including predicting a student's ability to succeed on a large-scale summative assessment...diagnosing gaps in a student's learning. Many of the assessments currently in use that are labeled "benchmark," "formative," "diagnostic," or "predictive" fall within our definition of interim assessments. (pp.1-2)

By setting testing windows for the pretests and posttest at the district level and by using the data to make predictions regarding NJ ASK student achievement, the DE District applied FAT more as an interim assessment rather than a formative assessment.

In the state of New Jersey, FAT provides participating school districts with National Benchmark assessments. According to The Company (2009) "National Benchmarks provide snapshots of student achievement so that administrators and teachers can target students early who need intervention or additional resources" (p.1). The Company explained that the National Benchmark assessments developed for FAT are done so in alignment to the state mandated accountability tests for each participating state. Within the school district and by FAT representatives, these National Benchmarks are more commonly referred to as the pretest and posttest, or Test A and Test B.

Central office administrators established mandates as to when pre and post testing opportunities would be permitted for the district included in this study. FAT representatives suggested that Test A be administered in the beginning of the academic school year and Test B be administered some time prior to the administration of the NJ ASK. In essence, the FAT representatives recommended using the product as interim assessments. In an effort to target students who need interventions and to drive classroom instruction, the District opted to administer Test B in March, 1 month prior to the state administration of the NJ ASK in hopes of identifying those students who were close to proficiency and then provide a surge of test preparation to those students. Booher-Jennings (2005) described this practice as educational triage for the "bubble kids."

FAT also provides limited item banks of math and reading questions that require teachers to design, develop, and assign individualized assessments to "evaluate instructional effectiveness." FAT does not provide pre-built assessments to use as interim assessments, instead it is the responsibility of teachers, administrators, and/or curriculum leaders, to develop tests using the item banks furnished by FAT or by importing their own generated tests. The Company does not suggest how often districts should take advantage of the item bank sample questions to generate their own tests, but instead recommends that each district uses it at their discretion.

The Company reports that all other tests besides Test A and Test B created using FAT are "interim administered assessments" and should be implemented "during an instructional block to measure student progress relative to grade-level learning objectives" (p.1). In essence, The Company recommends using FAT as both a formative and interim assessment tool. This is problematic because (a) FAT as defined by The Company is a "Formative Assessment System" and is sold/presented as such. and (b) that in order to take advantage of the formative assessment options, a great deal of additional support and training is required on the part of the district and subsequently on the teachers. The utility of such an approach would require a series of district mandates on how to ensure and monitor that all teachers are properly and effectively incorporating the formative assessments accordingly. However, it is important to note that the reviewed research regarding interim assessments indicated that there is no significant relationship between the use of interim assessments and informed instructional change in the classroom (Goertz, Olah, & Riggan, 2009).

The standardized betas and structure coefficients suggested that the FAT LA Pretest results have approximately the same power to predict student achievement on the language arts portion of the NJ ASK8 in all four schools as the FAT LA Posttest. In addition, the standardized betas and structure coefficients suggested that the FAT Math Pretest results have approximately the same predictive power in terms of standardized betas and structure coefficients as the FAT Math Posttest results in three of the four schools (School A, B, & D). For one school, School C, the strongest predictor of student achievement on the mathematics NJ ASK8 was the FAT Math Posttest.

If the predictive characteristics of both the pretests and the posttests are so similar in power for FAT language arts and mathematics for three of the four schools, and for language arts in all four schools, what instructional value is it to administer the pretest and the posttest— if the pretest can predict student achievement on the NJ ASK in September at the similar level as it does when administered in March? From this particular study it is impossible to determine to what degree and in what manner the teachers used the students' pre and posttests results to monitor and adjust their instruction. Therefore, it can be concluded that the mere act of pre and post testing in this sense may not be enough to create instructional change that manifests itself into increasing student achievement. Although it is unlikely that none of the teachers used the results from the FAT pretests to inform instruction, I was unable to determine how they actually used the results.

More of an effort must be made to ensure that formative assessment products marketed to schools and formative assessment practices used in schools are indicative of the type found most effective represented in the literature. That is, the most effective formative assessment occurs frequently and it is structured to provide opportunities for the students to practice the technique of self-evaluation, and reflection in order to monitor and adjust their learning. An experimental study conducted by Schunk (1996) revealed that students who had structured opportunities to practice the formative assessment technique of self-evaluation frequently, had greater motivation and increased achievement outcomes in comparison to those who did not participate in the practice of self-evaluation frequently.

The existing empirical literature and the results from this study seem to suggest that the more proximal (closer to the student) the formative assessment activity is (i.e., self-evaluation), the greater the influence it has on learning (Sadler, 1989; Schunk, 1996) whereas the further the formative assessment is from the student (distal), the less influence it has on learning. The "formative" assessment product, FAT is distal from the student as used in this district and as recommended by the corporation. The student is not actively involved in self-monitoring or self-assessing. It is unclear how the product, as currently used in the four schools and marketed by the corporation facilitates reflection beyond superficial error identification.

Perie, Marion, and Gong (2007) and Heritage (2010) maintain formative assessment is a process employed to assist teachers and students with the necessary diagnostic feedback that would in turn reshape and drive classroom instruction. However, in order for a formative assessment activity to prove fruitful, Perie et al. emphasized "the assessment is embedded within the learning activity and linked directly to the current unit of instruction" (2007, p.1). Contrary to this definition, the FAT "formative activity" consisted of a language arts and mathematics pre and posttest administered at two specific, pre-determined intervals. In fact, the results from this study suggested that the pretests for three of the four schools in both language arts and mathematics were not formative.

If the testing product does not meet the definition found in the empirical literature of formative assessment, and the district leadership or teachers do not use it as such, then it is not a formative assessment. The product and the way it is used aligns with the definition of an interim assessment as found in the literature. The literature on interim assessment presented earlier indicates that at this time there is not a clear link between interim assessment practices and increased student achievement. The results from this study align to the literature.

As noted earlier, the assessment product used in the district and 177 other districts, will no longer be free to any New Jersey school districts beginning in the 2011-2012 school year. Due to substantial budget cuts incurred by the State, for the remaining year of the 5 year plan (2011-2012) and every year thereafter, FAT will only be available for a cost to each school district. The current rate to implement FAT is $6.38 per student. District leaders with limited resources who search for positive interventions will have to weigh the costs and the potential benefits of pre and post testing with this product or mere act of pre and post testing regardless of product.

Recommendations for Policy and Practice

Due to the recent FAT budget cuts and the early expiration of the free trial period, it is imperative that district administrators evaluate the effectiveness of FAT to justify the cost of $6.38 per student prior to purchasing this product. Although all the schools included in my study were located in the same DE school district, some differences among the schools in results did emerged. The results of my study raise the possibility that interventions geared toward increasing student achievement could be content and site (context) specific. The demographic characteristics of Grade 8 students for School C (n=200), 35% of whom were White, 13.5% Black, 16% Asian, 35% Hispanic, and .5% American Indian. Of the 200 students, 44.5% were classified as economically disadvantaged; this was the school with the greatest percentage of economically disadvantaged students in the district (see Table 36). The FAT Posttest results were only more predictive in math for School C compared to the other schools that had a lower percentage of students eligible for free or reduced lunch. However the language arts findings for School C were similar to the other three schools. It could be that the demographic make-up of a school (i.e., ethnicity and socioeconomic status) and the specific content, in this case, mathematics, influences which school variables influence student performance on the NJ ASK8. There is very little evidence for the predictive validity of commercial benchmark tests in predicting achievement on state tests (Brown & Coughlin, 2007). School leaders should examine the literature on the subject clearly and evaluate such products against the prevailing literature before bringing such interventions into the learning environment. Perhaps one-size fits all approach to applying interventions do not fit all.

Despite the fact that the New Jersey Department of Education and the United States Department of Education continue to set forth reform mandates that include the implementation and continued utility of formative assessment programs and practices, perhaps policies should not be universally prescribed to all school districts and all schools within a district. The findings of this study raise the possibility that there is no "one size fits all" model that is successful in predicting student achievement outcomes. If administrators and education leaders are searching for interventions that will work in multiple buildings within the same jurisdiction, these results provide empirical evidence, albeit a small amount, that this might not be possible with all interventions. What is successful in one building and location might not necessarily yield the same results in another, even within the same school district or DFG. Within the same district there can exist vast between-school differences in terms of demographics that influence achievement.

Table 36

Demographic Characteristics of Grade 8 Students by School with Percentages of Economically Disadvantaged

Ethnicity Breakdown									Total
School	Total	White	Black	Asian	Pacific Islander	Hispanic	Amer. Indian	Other	% Econ. Dis.
A	227	97	49	33	1	47	0	0	35%
B	229	165	19	22	0	22	1	0	13.5%
C	200	70	27	32	0	70	1	0	44.5%
D	252	97	29	99	0	27	0	0	23%
E	172	96	20	10	0	46	0	0	24%

The data from this study and others (Dusenbury et al. 2003; Glennan et al., 2004; Ruiz-Primo, 2005; Stein et al. 2008) suggests that at the very least, the federal and state agencies should consider that interventions act differently in different contexts and contents. Intervention results vary across schools due to several factors including "variation among teachers, schools, and fidelity of implementation and performance" (Stein et al., 2008, p.369). In a review of the literature Stein et al. (citing Glennan, Bodilly, Galeghar, & Kerr, 2004; Ruiz-Primo, 2005) discovered that two areas of "fidelity of implementation can vary (a) program characteristics and (b) features of settings into which the programs are placed" (p.370).

Conditions that influence the strength of implementation fidelity include the school organizations, teachers, and classrooms. Those factors are by no means equal in quality or context within a state, nation, or even a school district. As a result, varying conditions and contexts affect student academic achievement. When discussing interventions it is important to consider how individual student characteristics (e.g., gender, socioeconomic status, race/ethnicity, special education modifications, and English language learner status) also influence student achievement (Stein et al., 2008). The literature revealed that the social context of an intervention in regards to education setting and influential change professional development has upon teachers is also a relevant component (Dusenbury et al. 2003; Ruiz-Primo, 2005; Smylie, 1988; Stein et al. 2008). Stein et al. confirmed "that school-level characteristics are related to teacher fidelity of implementation of educational interventions and student achievement. These include (a) the instructional leadership of the principal and (b) school climate, staff morale, and communication within the school community" (p.373) (citing Dusenbury et al., 2003, Fullan & Pomfret, 1977, Gottfried, 1984).

Additional considerations for policy and practice include the use of raw scores rather than proficiency categories on assessments to make high-stakes decisions for students. Because proficiency categories are used in the district of study and at the state level to judge academic achievement, it is difficult to determine how much academic growth a student achieves from the FAT pre/posttest results because the variance in student scores is masked by a large, blunt, proficiency category. The same problem exists for the NJ ASK. The state judges districts on the percentage of students rated as proficient or higher, not on how much scale scores increased or decreased. There can be large increases in scale scores but no increase or even a drop in the percentage of students rated as proficient. States and districts should at the very least use continuous variables as part of their data set to make judgments about academic achievement. The continuous data (scale scores) allow you to see the growth, whereas proficiency categories mask much of variance and make it much harder to determine growth.

Academic Support Instruction Variable

Conclusions

Three of the four schools included in the study offered Academic Support Instruction (ASI) services, School A was the school that did not. The results of the study revealed that for the three schools that did provided ASI services, Schools B, C, and D, ASI was a significant negative predictor of student achievement for both language arts and mathematic performance on the NJ ASK8. Included in Chapter III was the criteria used by the DE District to provide ASI services to students. Students were enrolled in ASI courses if they scored below the designated levels (200) of proficiency on the NJ ASK7 for language arts and mathematics. Therefore, the results of the study regarding the negative relationship ASI services has with student achievement is consistent with the literature and should come of no surprise. Students' future academic achievements are influenced and can be predicted by past academic achievements (Adelman, 2006; Dossett & Munoz, 2000; Ingels et al., 2002; Smith, 2006;).

Recommendations for Programs

School administrators should conduct impact studies of their ASI programs given that ASI is related to academic achievement on the NJ ASK. Well constructed impact studies use control groups and/or matched pairs in their designs and make systematic attempts to control for confounding variables (Song, 2010). The impact study can provide school leaders with information needed to determine program effectiveness.

Attendance Variable

Conclusions

Attendance was a strong predictor of student achievement on the NJ ASK8 for language arts and mathematics for two of the four schools included in the study, Schools A and B. In School D, attendance was identified as a significant predictor for mathematics achievement only. In Schools A and B poor attendance, defined as exceeding the district's attendance policy of 16 absences, influenced student achievement negatively. It is important to take note of the demographic characteristics of Grade 8 students in School B. The demographic breakdown in School B (n=229) was as follows: 72% of students were White, 8.3% Black, 9.6% Asian, 9.6% Hispanic, and .5% American Indian, 13.5% were classified as economically disadvantaged; this was the school with the lowest percentages of economically disadvantaged students in the district (see Table 36). These findings suggest that perhaps a particular school's SES is a suppressor variable. When SES is not a concern (e.g., low free and reduced eligibility percentages), the variable of attendance emerges as a significant predictor of student achievement because it is not being suppressed by SES. As a suppressor variable SES could be masking interactions between it and other variables, and as a result produced the findings discussed here. Although it was found for both subjects in two schools and only mathematics in another, the findings that poor attendance is associated with poor performance on the NJ ASK8 is in alignment with the literature. Three studies in particular, those conducted by Caldas (1993), Roby (2004), and Sheldon (2007) confirmed that student attendance has a statistically significant relationship with student achievement on standardized tests.

Recommendations for Policy, Practice, and Programs

Based on the findings of this study and review of the literature, it is important that school administrators take a proactive approach toward student absenteeism. District leaders need to put in place measures that prevent students from falling behind due to excessive absences. For example, rather than wait for students to reach the threshold of days absent prior to being in violation of the attendance code, intervention and prevention programs could be established that require action be taken for every third student absence. It is also important to enforce the New Jersey state code that requires home instruction services be provided for students that miss school for five consecutive days, regardless of whether the absences are due to suspension, illness, or for personal reasons. Effort needs to be made on the part of school leaders to ensure that early interventions are in place for students that are absently frequently prior to the point that it begins to affect student achievement.

Gender Variable

Conclusions

Gender significantly predicted student achievement on the NJ ASK8 for language arts in three of the four schools (Schools A, B, and D). Gender significantly predicted student achievement on the NJ ASK8 for mathematics in only one of the four schools, School C. In Schools A, B, and D the betas and structure coefficients suggested that being female (in addition to other variables) predicted a percentage of the variance in student performance on the NJ ASK8 language arts test. For School C, the structure coefficient confirmed that gender was a student variable that predicted student achievement. Male students in School C tended to perform poorer on the NJ ASK8 mathematics test. Gender was not a significant variable in predicting mathematics performance in three of the four schools (Schools A, B, and D).

The results of the study do not reflect what was reviewed in the literature regarding gender differences. It appears from the results of the study that females may perform better on the NJ ASK8 language arts assessment. A review of the literature revealed that although the stereotype is that females outperform males in liberal arts and that males outperform females in mathematics and the sciences, there is little truth to this. Research shows that although there was once a gender gap in mathematics and sciences they have since diminished. This is evident in all the schools except for School C. There is little empirical evidence concerning the gender gap in liberal arts today. One study reviewed did find that data from 1990 confirmed that girls had a slight advantage in writing over boys. The results of this study could have been attributed to the curriculum, the demographic characteristics of the schools, or the particular passages and tasks selected for the NJ ASK8 assessment.

Recommendations for Policy and Practice

Due to the recent limited empirical evidence regarding gender differences in language arts and mathematics achievement great caution should be exercised when considering the predictability this variable has on student performance. A more proactive approach an administrator could take is to regularly monitor the achievement gap (if one exists) between gender for language arts and mathematics to assure there are no red flags raised year to year. It is important to ensure that both genders have equal opportunities to interact and engage with all subject content areas throughout a student's education career, regardless of the former stereotypes that mathematics and sciences appeal more to males and the liberal arts generally to females. By vigilantly monitoring data for trends, administers will easily be able to identify a problem at the first sign of one, and immediately put in place the necessary actions to rectify the situation.

Socioeconomic Status Variable

Conclusions

Coleman et al. reported in 1966 that the greatest influence on student academic performance was socioeconomic status (SES), followed by teacher characteristics and class size. Over 40 years after the release of the Coleman Report (1966), much of the reviewed literature continues to support the original findings of Coleman et al. After reviewing the extensive literature available regarding the potential attainment of educational equality among students it is evident that enacting accountability policies, providing additional funding, using high-stake consequences and the results from those tests as major indicators of student academic success, and providing an increased number of education resources to struggling schools will not, in and of themselves, lead to the successful bridging of existing achievement gaps at the state and national testing level (Lee &Wong, 2004). The findings of this study support that conclusion.

When analyzing betas it was discovered that in two of the four schools SES was a statistically significant and strong predictor of student achievement for both language arts and mathematic performance on the NJ ASK8 in Schools B and C in terms of interpreting the betas. The demographic breakdown in School B (n=229) was as follows: 72% of students were White, 8.3% Black, 9.6% Asian, 9.6% Hispanic, and .5% American Indian, 13.5% were classified as economically disadvantaged; this was the school with the lowest percentages of economically disadvantaged students in the district (see Table 36). The demographic characteristics of Grade 8 students for School C (n=200), 35% of whom were White, 13.5% Black, 16% Asian, 35% Hispanic, and .5% American Indian. Of the 200 students, 44.5% were classified as economically disadvantaged; this was the school with the greatest percentage of economically disadvantaged students in the district (see Table 36).

It has been discussed several times throughout this study that structure coefficients paint a more complete picture of the regression results and help to uncover potentially important predictors that would not have been given credit if the researcher analyzed only beta weights. This was in fact the case for School A and School D in this study. The SES beta weights for these two schools were not statistically significant and the analysis would have concluded there for many researchers. However, there is important information to report regarding the results of School A and School D related to the SES variable.

The results from School A suggest that SES might have acted as a suppressor variable in this case for language arts and mathematics student achievement with a small structure coefficient. SES has a long history of influencing student achievement. In a multiple regression model, SES also "drags" on other achievement variables that could relate to achievement on the NJ ASK. In other words, SES suppresses the predictive power of other variables and thus, itself is an indirect predictor of achievement for School A. When looking specifically at student variables that predicted language arts achievement SES emerged as the strongest, followed by gender then attendance. Similar findings were discovered for School D mathematics performance. The structure coefficients suggested that attendance, followed closely by SES were the strongest predictors of student achievement.

Therefore, the results of the study reflect what the literature suggests, that SES remains a strong predictor of student achievement. Regardless of a school's demographic characteristics, whether the school has the highest or lowest percentage of economically disadvantaged students in the district, the study confirmed that individual student's SES influenced their language arts and mathematics performance on the NJ ASK8.

Recommendations for Policy and Practice

The evidence collected from this study suggests that federal, state, and local agencies should reconsider the allocation of funds, especially excessively large amounts that aim to abolish the achievement gaps. The problem then becomes that the same high-stakes tests are used to track students. District leaders must seriously consider the ramifications associated with placing high-stakes decisions on one state mandated assessment. In some cases this one test will be the navigation course of a student's entire future learning experience.

Recommendations for Future Research

Although this study provided empirical evidence, it is not possible that a single explanatory study could provide all the answers to the multifaceted overarching research question of what student and school variables explain the greatest variance in student achievement as measured by the NJ ASK8 language arts and mathematics assessments. Therefore, it is important to conduct future research in the area of student and school variables that may influence student achievement.

1. For a researcher to use the actual FAT "percentage correct" scores (continuous variables) and the dichotomous variable of proficiency categorization in another analysis to determine if the addition of the percentages predict more variance in student achievement on the NJ ASK language arts and mathematics assessments.

2. Conduct a similar study using the NJ ASK6 and NJ ASK7 assessments to compare and contrast the findings of those to this study to investigate further the context specific nature of these variables.

3. Conduct this study at the student level and examine the influence of teacher characteristics on student achievement.

4. Conduct a similar study to determine if student perceptions of the Grade 8 language arts test influenced the gender gap reported in this study.

5. Conduct a similar study to analyze growth between FAT pre-post percentages and the relationship to NJ ASK scores.

6. Conduct a study to examine the relationship between student mobility (years a student has been in the district) and NJ ASK scores.

7. Identify schools whose Grade 8 students are outperforming other Grade 8 students in the same DFG with similar student and school variables, such as FAT use. One can further investigate if there were any measureable differences in which FAT, or a similar commercially produced formative/interim assessment tool was used to successfully improve student learning.

The results of this study, the FAT formative/interim assessment misclassification, and the continued use of similar commercially produced products in conjunction with other student variables that cannot be altered (i.e., student's gender, SES, attendance record) and the continued use of test results from state mandated assessments for high-stakes decisions, suggest that further study on the influence of these student and school variables is warranted.

References

Adams, R.D., Hutchinson, S. & Martray, C. (1980). *A developmental study of teacher concerns across time.* Paper presented at the annual meeting of the American Educational Research Association. Boston, Mass.

Adelman, C. (2006). *The Toolbox revisited: Paths to degree completion from high school through college.* Washington, DC: U.S. Department of Education, Office of Vocational and Adult Education.

Aikin, W. (1942). *The Story of the Eight-Year Study.* New York: Harper.

Amaker, N.C. (1988). *Civil rights and the Reagan Administration.* Washington, D.C.: The Urban Institute Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for education and psychological testing.* Washington, D.C.: American Educational Research Association, 142.

American Productivity and Quality Center (APQC). (2005). *Benchmarking to improve an examination of districts' processes in: Assessment.* Retrieved August 16, 2009, from http://www.apqc.org/educ/docs/Final%20Assessment%20Mock%20Report_Complete.pdf

Amrein, A.L., & Berliner, D.C. (2002a). *The impact of high-stakes tests on student academic performance: An analysis of NAEP results in states with high-stakes tests and ACT, SAT, and AP Test results in states with high school graduation exams.* Retrieved January 13, 2010, from Education Policy Studies Laboratory, Education Policy Research Unit: http://edpolicylab.org

Amrein, A.L., & Berliner, D.C. (2002b). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10(*18). Retrieved January 13, 2010, from http://epaa.asa.edu/epaa/v10n18/

Amrein, A.L., & Berliner, D.C. (2003). The effects of high-stakes testing on student motivation and learning. *Educational Leadership*, *60*(5), 32. Retrieved January 13, 2010, from Academic Search Premier database.

Amrein-Beardsley, A.L., & Berliner, D.C. (2003, August). Re-analysis of NAEP math and reading scores in states with and without high-stakes tests: Response to Rosenshine. *Education Policy Analysis Archives, 11*(25). Retrieved January 13, 2010, from http://epaa.asu.edu/epaa/v11n25/

Archbald, D., & Porter, A. (1990). *A retrospective and an analysis of roles of mandated testing in education reform.* Retrieved February 10, 2010, from ERIC database.

Association for Career and Technical Education (2006). Study on the impact of high-stakes testing. Retrieved January 18, 2010, from http://www.acteonline.org/content.aspx?id=4686&terms=Techniques-+January+2006

Averch, H. A., Carroll, S. J., Donaldson, T. S., Kiesling, H. J., & Pincus, J. (1974). *How effective is schooling? A critical review of research.* Englewood Cliffs, NJ: Educational Technology Publications.

Baker, K. (1991). Yes, throw money at schools. *Phi Delta Kappan, 72*(8), 628-631.

Baker, E. L., & Linn, R. L. (2002). *Validity issues for accountability systems.* Center for the Study of Evaluation. Technical Report 585, Los Angeles, CA.

Barnes, S., Salmon, J., &Wale, W. (1989, March). *Alternative teacher certification in Texas*. Presented at the annual meeting of the American Educational Research Association, San Francisco, CA. (ERIC Document No. 307316).

Bell, B., & Cowie, B. (2000). The characteristics of formative assessment in science education. *Science Education, 85*, 536–553.

Bernauer, J.A., & Cress, K. (1997). How school communities can help redefine accountability assessment . *Phi Delta Kappan, 79*(1), 71-75.

Bergan, J.R., Sladeczek, I. E., Schwarz, R. D., & Smith, A. N.(1991). Effects of a measurement and planning system on kindergartners' cognitive development and educational programming. *American Educational Research Journal, 28*, 683-714.

Betts, J., Rice, L. & Zau, J. (2003). *Determinants to student achievement: New evidence from San Diego*. San Francisco, CA: Public Policy Institute of California.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education,* 5(1), 7-74.

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal, 42*(2), 231-268.

Boote D. N., & Beile, P. (2005). Scholars before researchers: On the centrality of the dissertation literature in research preparation. *Educational Researcher, 34*(6), 3-15.

Bowles, S., & Levin, H. M. (1968). The determinants of scholastic achievement—An appraisal of some recent evidence. *The Journal of Human Resources, 3*(1), 3–24.

Boyd, D., Goldhaber, D., Lankford, H., & Wyckoff, J. (2007). The effect of certification and preparation on teacher quality. *Future of Children*, *17*(1), 45-68. Retrieved February 10, 2010, from ERIC database.

Boyle, E., Duffy, T., & Dunleavy, K. (2003). Learning styles and academic outcome: The validity and utility of Vermunt's Inventory of Learning Styles in a British higher education setting. *British Journal of Educational Psychology*, *73*(2), 267-290. Retrieved November 8, 2009, from Academic Search Premier database.

Braun, H. (2004, January 5). Reconsidering the impact of high-stakes testing. *Education Policy Analysis Archives, 12*(1). Retrieved January 13, 2010, from http://epaa.asu.edu/epaa/v12n1/

Brooks, T., & Pakes, S. (1993). Policy, national testing, and the psychological corporation. *Measurement & Evaluation in Counseling & Development (American Counseling Association)*, *26*(1), 54. Retrieved February 13, 2010, from Academic Search Premier database.

Brown, R.S. & E. Coughlin. (2007). *The predictive validity of selected benchmark assessments used in the Mid-Atlantic Region* (Issues & Answers Report, REL-2007-No. 017). Washington DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Educational Laboratory Mid-Atlantic. Retrieved December 1, 2010, from http://ies.ed.gov/ncee/edlabs

Buchanan, T. (2000). The efficacy of a World-Wide Web mediated formative assessment. *Journal of Computer Assisted Learning, 16*, 193-200.

Butler, R. (1988). Enhancing and undermining intrinsic motivation; the effects of task-involving and ego-involving evaluation on interest and performance. *British Journal of Educational Psychology, 58*, 1-14.

Caldas, S. J. (1993). Reexamination of input and process factor effects in public school achievement. *The Journal of Educational Research, 86*(4), 206-214.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis, 24*(4), 305-331.

Chen, C., & Stevenson, H. W., (1995). Motivation and mathematics achievement: A comparative study of Asian-American, Caucasian, and East Asian high school students. *Child Development, 66*(4), 1215-1234.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (2nd ed). New York: Academic Press.

College Board (2010). *SAT reasoning test. What is the SAT?* Retrieved January 13, 2010, from http://professionals.collegeboard.com/testing/sat-reasoning

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfield, F.D., & York, R.L. (1966). *Equality of educational opportunity.* Washington, DC: U.S. Government Printing Office.

Courville, T., & Thompson, B. (2001). Use of structure coefficients in published multiple regression articles. Beta is not enough. *Educational and Psychological Measurement, 61,* 229-248.

Creswell, J.W. (1994). *Research design: Qualitative and quantitative approaches.* Thousand Oaks, CA: Sage.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*(4), 438-481.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8*(1). Retrieved January 13, 2010, from http://epaa.asu.edu/epaa/v8nl

Darling-Hammond, L. (2004). From "separate but equal" to "No Child Left Behind": The collision of new standards and old inequalities. In D. Meier & G. Woods (Eds.), *Many children left behind: How the No Child Left Behind Act is damaging our children and our schools* (pp.3-32). Boston: Beacon Press.

Darling-Hammond, L., Berry, B. & Thoreson, A. (2001). Does teacher certification matter? Evaluating the evidence. *Education Evaluation and Policy Analysis, 23*(1), 57-77.

Darling-Hammond, L., Hudson, L., & Kirby, S. (1989). *Redesigning teacher education: Opening the door for new recruits to science and mathematics teaching.* Santa Monica: RAND.

Davy, L. (2008). *Learnia formative assessment resources 2008-2009.* Retrieved August 15, 2009, from http://www.state.nj.us/education/assessment/formative/memo050908.pdf

Denton, J.J., & Lacina, L.J. (1984). Quantity of professional education coursework linked with process measures of student teaching. *Teacher Education and Practice*, 39-64.

Dorn, S. (1998). The political legacy of school accountability systems. *Education Policy Analysis Archives*, *6*(1), 1-31. Retrieved January 13, 2010, from http://epaa.asu.edu/epaa/v6n1

Dossett, D., & Munoz, M. A. (2000). *Educational reform in the accountability era: The impact of prior achievement and socio-economic conditions on academic performance.* (ERIC Document Reproduction Services No. 468 491)

Drukker, M., Kaplan, C., Schneiders J., Feron, F.J., & van Os, J. (2009). The wider social environment and changes in self-reported quality of life in the transition from late childhood to early adolescence: A cohort study. *BMC Public Health, 6,* 133.

Dryfoos, J.G. (1996). Adolescents at risk: Shaping programs to fit the need. *The Journal of Negro Education, 65*(1), 5-18.

Dunn, K.E., & Mulvenon, S.W. (2009a). A critical review of research on formative assessments: The limited scientific evidence of the impact of formative assessments in education. *Practical Assessment, Research & Evaluation,* 14(7), 1-11. Retrieved July 27, 2009, from http://pareonline.net/pdf/v14n7.pdf

Dunn, K.E., & Mulvenon, S.W. (2009b). *Let's talk formative assessment…and evaluation?* Retrieved August 16, 2009, from http://eric.ed.gov (ERIC Document Reproduction Service No. ED505357)

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research, 18*(2), 237-256.

Edelman, M. (1997). Leaving no child behind. *School Administrator, 54,* 14-16.

Elementary and Secondary Act of 1965 (P.L. 89-10).

Else-Quest, N.M., Hyde, J.S., & Linn, M.C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 10*(1), 103-127.

Evertson, C., Hawley, W., & Zlotnick, M. (1985). Making a difference in educational quality through teacher education. *Journal of Teacher Education, 36*(3), 2-12.

Feiman-Nemser, S., & Parker, M. B.(1990). Making Subject Matter Part of the Conversation or Helping Beginning Teachers Learn to Teach. East Lansing, MI:National Center for Research on Teacher Education.

Ferguson, R. F.(1991). Paying for public education: New evidence on how and why money matters. *Harvard Journal on Legislation, 28*(2), 465-498.

Ferguson, R. (1998). Teachers' perceptions and expectations and the Black-White test score gap. In C. Jencks and M. Phillips (Eds.), *The Black- White test score gap* (pp. 273-317). Washington, DC: Brookings Institution Press.

Ferguson, R. F., & Womack, S. T. (1993). The impact of subject matter and education coursework on teaching performance. *Journal of Teacher Education, 44*(1), 55-63.

Fetler, M. (1999). High school staff characteristics and mathematics test results. *Education Policy Analysis Archives, 7*(9). Retrieved February 10, 2010, from http://epaa.asu.edu/epaa/v7n9/

Finn, J. D. (1993). *School engagement and students at risk.* Washington, DC: National Center for Education Statistics.

Finn, J. P., & Achilles, C. M. (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal, 27*, 557-577.

Finn, J. P., & Achilles, C. M. (1999). Tennessee's class size study: Findings, implications, and misconceptions. *Educational Evaluation and Policy Analysis, 21,* 97–110.

Fontana, D., & Fernandes, M. (1994). Improvements in mathematics performance as a consequence of self-assessment in Portuguese primary school pupils. *British Journal of Educational Psychology, 64*, 407-417.

Fuchs, L.S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children, 53,* 199-208.

Gamoran, A., & Long, D. A. (2006). *Equality of educational opportunity: A 40-year retrospective* (WCER Working Paper No. 2006-9). Madison: University of Wisconsin–Madison, Wisconsin Center for Education Research. Retrieved February 12, 2010, from http://www.wcer.wisc.edu/publications/workingPapers/papers.php

Garan, E.M. (2004). *In defense of our children: When politics, profit, and education collide.* Portsmouth, NH: Heinemann.

Garson, D. (2010). *Multiple regression.* Retrieved April 13, 2010, from http://faculty.chass.ncsu.edu/garson/PA765/regress.htm

Glass, G.V., Cahen, L.S., Smith, M.L., & Filby, N.N. (1982). *School class size: Research and policy.* Beverly Hills, CA: SAGE Publications.

Glassberg, S. (1980). *A view of the beginning teacher from a developmental perspective.* Paper presented at the American Educational Research Association Annual Meeting. Boston, MA.

Glennan, T. K., Bodilly, S. J., Galegher, J. R., & Kerr, K. A. (Eds.). (2004). *Expanding the reach of education reforms: Perspectives form leaders in the scale-up of educational interventions.* Santa Monica, CA: RAND.

Goals 2000: Educate America Act (P.L. 103-227).

Goebel, S. D., Romacher, K., & Sanchez, K. S. (1989). *An evaluation of HISD 's alternative certification program of the academic year: 1988-1989.* Houston, TX: Houston Independent School District Department of Research and Evaluation. (ERIC Document No. 322103)

Goertz, M. E., Olah, L. N., & Riggan, M. (December, 2009). *Can interim assessments be used for instructional change?* Consortium for Policy Research in Education Policy Brief.

Goldhaber, D. (2002). *The mystery of good teaching*. Retrieved January 13, 2010, from http://educationnext.org

Goldhaber, D. & Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis, 22*(2), 129-145.

Goldschmidt, P., & Wang, J. (1999). When can schools affect dropout behavior? A longitudinal multilevel analysis. *American Educational Research Journal, 36*(4), 715-738.

Gomez, D. L., & Grobe, R. P. (1990). *Three years of alternative certification in Dallas: Where are we?* Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Good, T.L., & Brophy, J.E. (1995). *Contemporary educational psychology.* Reading, MA: Addison, Wesley, and Longman.

Gottfried, M. A. (2009). Excused versus unexcused: How student absences in elementary school affect academic achievement. *Educational Evaluation and Policy Analysis, 31*(4), 392-415.

Greenwald, R., Hedges, L.V., & Laine, R.D. (1996a). The effect of school resources on student achievement. *Review of Educational Research, 66,* 361-396.

Greenwald, R., Hedges, L.V., & Laine, R.D. (1996b). Interpreting research on school resources and student achievement: A rejoinder to Hanushek. *Review of Educational Research, 66,* 441-416.

Grossman, P. L. (1989). Learning to teach without teacher education. *Teachers College Record, 91*(2), 191-208.

Gunzelmann, B., & Connell, D. (2006). The new gender gap: Social, psychological, neuro-biological and educational perspectives. *Educational Horizons, 84*(2), 94-101.

Gurian, M., & Stevens, K. (2004). With boys and girls in mind. *Educational Leadership*, 62(3), 21-26.

Guyton, E., & Farokhi, E. (1987, Sept-Oct). Relationships among academic performance, basic skills, subject matter knowledge and teaching skills of teacher education graduates. *Journal of Teacher Education, 38,* 37-42.

Hanushek, E.A. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *The American Economic Review, 61*(2), 280-288.

Hanushek, E.A. (1972). *Education and race: An analysis of the educational production process.* Lexington, MA: D.C. Heath.

Hanushek, E. A. (1979). Conceptual and empirical issues in the estimation of educational production functions *The Journal of Human Resources, 14*, 351–388.

Hanushek, E.A. (1981). Throwing money at schools. *Journal of Policy Analysis and Management, 1,* 19-41.

Hanushek, E.A.(1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature, 24,* 1141-1177.

Hanushek, E.A., (1989). The impact of differential expenditures on school performance. *Educational Researcher, 18*(4), 45-65.

Hanushek, E. A. (1991). When school finance "reform" may not be good policy. *Harvard Journal on Legislation, 28,* 423-456.

Hanushek, E. A. (1994). *Making schools work: Improving performance and controlling costs.* Washington, DC: Brookings Institution.

Hanushek, E. A. (1996). A more complete picture of school resource policies. *Review of Educational Research, 66*, 367–409.

Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis, 19,* 141–164.

Hanushek, E. A., & Kain, J. F. (1972). On the value of equality of educational opportunity as a guide to public policy. In F. Mosteller & D. P. Moynihan (Eds.), *On equality of educational opportunity* (pp.116–145). New York: Vintage Books.

Hawk, P., Coble, C.R., & Swanson, M. (1985). Certification: It does matter. *Journal of Teacher Education, 36*(3), 13-15.

Hawkins, E. F., Stancavage, F. B., & Dorsey, J. A. (1998). *School policies affecting instruction in mathematics.* Washington, D.C.: National Center for Education Statistics.

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269*, 41-45.

Hedges, L.V., & Olkin, I. (1980). Vote counting method in research synthesis. *Psychological Bulletin, 88,* 359-369.

Henly, D.C. (2003). Use of web-based formative assessment to support student learning in a metabolism/nutrition unit. *European Journal of Dental Education, 7,* 116-122.

Heritage, M. (2010). *Formative assessment and next generation assessment systems: Are we losing an opportunity?* Washington, D.C. : Council of Chief State School Officers. Retrieved December 12, 2010, from http://cse.ucla/edu/products/misc/Formative_Assessment_Next_Generation_Heritage.pdf

Herman, J., & Golan, S. (1993). Effects of standardized testing on teaching and schools. *Educational Measurement: Issues and Practice, 12*(4), 20-25.

Hoenack, S.A., & Collins, E.L. (1990). *The economics of American universities: Management, operations, and fiscal environment.* Albany, NY: State University of New York Press.

Hong, W.P., & Youngs, P. (2008). Does high-stakes testing increase cultural capital among low-income and racial minority students? *Education Policy Analysis Archives, 16*(6). Retrieved February 12, 2010, from http://epaa.asu.edu/epaa/v16n6/

Howell, D. (2002). *Multiple regression #3*. Retrieved March 3, 2010, from http://www.uvm.edu/~dhowell/gradstat/psych341/lectures/MultipleRegression/multreg3.html

Hyde, J. S., Fennema, E., & Lamon, S. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107*(2), 139-155.

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A., & Williams, C. (2008). Gender similarities characterize math performance. *Science, 321*, 494-495.

Ingels, S. L., Curtin, T. R., Owings, J. A., Kaufman, P., Alt, M. N., & Chen, X. (2002). *Coming of age in the 1990s: The eighth-grade class of 1988 12 years later.* Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Jelmberg, J. (1996). College-based teacher education versus state-sponsored alternative programs. *Journal of Teacher Education*, *47*(1), 60-66. Retrieved February 13, 2010, from ERIC database.

Jencks, C., & Phillips, M. (1998). *The black-white test score gap: An introduction.* In C. Jencks & M. Phillips (Eds.), *The black-white test score gap* (pp. 1–51). Washington, DC: Brookings Institution Press.

Jencks, C., Smith, M., Acland, H., Bane, M. S., Cohen, D., Gintis, H., et al. (1972). *Inequality: A reassessment of the effect of family and schooling in America*. New York: Basic Books.

Jepsen, C., & Rivkin, S. (2009). Class size reduction and student achievement: The potential tradeoff between teacher quality and class size. *Journal of Human Resources, 44*(1), 223-250.

Johnson, B. (2001). Toward a New classification of nonexperimental quantitative research. *Educational Researcher*, *30*(2), 3-13.

Kennedy, E. (2003). *Raising test scores for all students: An administrator's guide to improving standardized test performance.* Thousand Oaks, CA: Corwin Press.

Kline, T.J.B. (2005). *Psychological testing: A practical approach to design and evaluation.* Thousand Oaks, CA: Sage Publications, Inc.

Klitgaard, R.E. & Hall, G.R. (1974). Are there unusually effective schools? *Journal of Human Resources, 10*(3), 90-106.

Kohn, A. (1993). *Punished by rewards: The trouble with gold stars, incentive plans, A's, praise, and other bribes.* New York: Houghton Mifflin.

Kohn, A. (2000). *The case against standardized testing: Raising the scores, ruining the schools*. Portsmouth, NH: Heinemann.

Laczko-Kerr, I., & Berliner, D.C. (2002, September). The effectiveness of "Teach for America" and other under-certified teachers on student academic achievement: A case of harmful public policy. *Education Policy Analysis Archives, 10(*37). Retrieved February 10, 2010, from http://epaa.asu.edu/epaa/v10n37/

Lee, J., & Wong, K. (2004). The Impact of accountability on racial and socioeconomic equity: considering both school resources and achievement outcomes. *American Educational Research Journal, 41*(4), 797-832.

Leech, N.L., Barrett, K.C., & Morgan, G.A. (2008). *SPSS for intermediate statistics: Use and interpretation* (3rd). New York, NY: Taylor & Francis Group.

Lehr, C. A., Sinclar, M. F., & Christenson, S. L. (2004). Addressing student engagement and truancy prevention during the elementary school years: A replication study of the Check & Connect model. *Journal of Education for Students Placed at Risk, 9*(3), 279-301.

Leung, C., & Mohan, B. (2004). Teacher formative assessment and talk in classroom contexts: Assessment as discourse and assessment of discourse. *Language Testing, 21*(3), 335-359.

Link, C.R., & Ratledge, E.C. (1979). Student perceptions, I.Q. and achievement. *Journal of Human Resources, 14*(1), 98-111. Retrieved February 10, 2010, from ERIC database.

Lopez, O. S. (1995). *Classroom diversification: An alternative paradigm for research in educational productivity.* Unpublished doctoral dissertation, University of Texas, Austin.

Madaus, G. (1988). The influence of testing on the curriculum. In L. Tanner 9ed.), *Critical issues in curriculum: 87th Yearbook of the NSSE Part 1.* Chicago: University of Chicago Press (ERIC Document Reproduction Service No. 263 183).

Marchant, G. (2004). What is at stake with high stakes testing? A discussion of issues and research. *Ohio Journal of Science, 104*(2), 2-7. Retrieved January 3, 2010, from Academic Search Premier database.

Marchant, G., Paulson, S., & Shunk, A. (2006). Relationships between high-stakes testing policies and student achievement after controlling for demographic factors in aggregated data. *Education Policy Analysis Archives*, *14*(30), 1-34. Retrieved January 3, 2010, from ERIC database.

Marks, G. N. (2008). Accounting for the gender gaps in student performance in reading and mathematics: Evidence from 31 countries. *Oxford Review of Education, 34*(1), 89-109.

Martinez, J.G. R., & Martinez, N. C. (1992). Re-examining repeated testing and teacher effects in a remedial mathematics course. *British Journal of Educational Psychology, 62*, 356-363.

Messick, S (1995). Standards-based score interpretation: Establishing valid grounds for valid interpretations. *Proceedings on the joint conference of standard setting for large-scale assessments,* Sponsored by the National Assessment Governing Board and the National Center for Educational Statistics. Washington, DC: Government Printing Office.

Michel, A.P. (2004). *What is the relative influence of teacher educational attainment on student NJASK4 scores?* Unpublished Doctoral Dissertation. Seton Hall University, South Orange, NJ.

Michel, A.P. (2008). Variables from the New Jersey school report card that predict student achievement on the NJASK4. *New Jersey Journal of Supervision and Curriculum Development,* 52, 34-45.

Mitchell, N. (1987). *Interim evaluation report of the alternative certification program* (REA87-027-2). Dallas, TX: DISD Department of Planning, Evaluation, and Testing.

Monk, D. (1994). Subject area preparation of secondary mathematics and science teachers and student achievement. *Economics of Educational Review, 12*(2), 125-142.

Mosby's Medical Dictionary (8th ed.) (2009). St. Louis: Elsevier.

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The Future of Children, 5*(2), 113- 127.

Mosteller, F., & Moynihan, D. P. (1972). A pathbreaking report: Further studies of the Coleman Report. In F. Mosteller & D. P. Moynihan (Eds.), *On equality of educational opportunity* (pp.3–68). New York: Vintage Books.

Murnane, R.J. (1975). *The impact of school resources on the learning of inner city children.* Cambridge, MA: Balinger Publishing Company.

Murnane, R.J., & Phillips, B.R. (1981). Learning by doing, vintage, and selection: Three pieces of the puzzle relating teaching experience and teaching performance. *Economics of Education Review,* 4, 691-693.

National Center for Education Statistics (2010). *NAEP technical documentation* . Retrieved March 1, 2010, from http://nces.ed.gov/nationsreportcard/tdw/scoring/

Natriello, G. (1987). The impact of evaluation processes on students. *Educational Psychologist, 22*(2), 155–175.

New Jersey Department of Education (2006a). *District factor groups for school districts.* Retrieved August 10, 2009, from http://www.state.nj.us/education/finance/sf/dfgdesc.shtml

New Jersey Department of Education. (2006b). *New Jersey assessment of knowledge and skills.* Retrieved August 25, 2009, from http://www.state.nj.us/education/assessment/

New Jersey Department of Education (2006c). *New Jersey core curriculum content standards.* Retrieved November 8, 2009, from http://www.state.nj.us/education/cccs/

New Jersey Department of Education (2009). *New Jersey Assessment of Skills and Knowledge 2008 technical report grades 5-8.* Retrieved July 28, 2009 from http://www.state.nj.us/education/assessment/ms/5-8/tech/2008TechReport.pdf

New Jersey Department of Education. (2010a). *Executive Order No.14.* Retrieved March 5, 2010, from http://www.state.nj.us/education/

New Jersey Department of Education. (2010b). *New Jersey Department of Education releases 2009 school report cards.* Retrieved March 5, 2010, from http://www.state.nj.us/education/news/2010/0209rc.htm

New Jersey Department of Education. (2010c). *Understanding the American Recovery and Reinvestment Act.* Retrieved February 25, 2010, from http://www.state.nj.us/recovery/index.shtml

New Jersey Department of Treasury (2001). *Local government budget review*. Retrieved August 10, 2009, from http://www.state.nj.us/treasury/lgbr/index.html

Nichols, S. L., Glass, G. V., & Berliner, D. C. (2006). High-stakes testing and student achievement: Does accountability pressure increase student learning? *Education Policy Analysis Archives, 14*(1). Retrieved January 3, 2010 from, http://epaa.asu.edu/epaa/v14n1

No Child Left Behind Act (2001). *PL107-110*. Washington, DC: Unites States Department of Education.

No Child Left Behind: A Desktop Reference. (September 2002). Retrieved February 25, 2010, from http://www.ed.gov/admins/lead/account/nclbreference/reference.pdf

OECD (2001). *Knowledge and skills for life. First results from the OECD programme for international student assessment* (Paris, Organization for Economic Co-operation and Development.

Pajas, F. (2002). Gender and perceived self-efficacy in self-regulated learning. *Theory Into Practice, 41*(2), 116-225.

Paulson, S.E., & Marchant, G.J. (2009). Background variables, levels of aggregation, and standardized test scores. *Education Policy Analysis Archives, 17*(22). Retrieved January 13, 2010 from, http://epaa.asu.edu/epaa/v17n22/

Pearson Education, Inc. (2009). *Assessment and information: Learnia*. Retrieved August 10, 2009, from http://pearsonassess.com/

Pedhazur, E.J. (1997). *Multiple regression in behavioral research: Explanation and prediction*. Fort Worth, TX: Harcourt Brace College Publishers.

Pedulla, J., Abrams, L., Madaus, G., Russell, M., Ramos, M., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers.* Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Perie, M., Marion, S., & Gong, B. (2007). *The role of interim assessments in a comprehensive assessment system: A policy brief.* Retrieved August 16, 2009, from http://www.achieve.org/files/TheRoleofInterimAssessments.pdf

Perie, M., Moran, R., & Lutkus, A. D. (2005). *NAEP 2004 trends in academic progress: Three decades of student performance in reading and mathematics* (NCES 2005–464). Washington, DC: Government Printing Office.

Piton Foundation & Dunnell-Kay Foundation (2007). *Testing in Colorado: Time, cost and purpose.* Retrieved August 10, 2009, from http://www.piton.org/Documents/PF-002%20Student%20Testing_r2.pdf

Plake, B.S. (2002). Evaluating the technical quality of educational tests used for high-stakes decisions. *Measurement and Evaluation in Counseling and Development,* 35(3), 144-152. Retrieved July 29, 2009, from ProQuest Psychology Journals. (Document ID: 264480721).

Popham, W.J. (1999). Why standardized tests don't measure educational quality. *Educational Leadership, 56*(6), 8. Retrieved from Academic Search Premier database.

Popham, W. J. (2001). *The truth about testing: An educator's call to action.* Alexandria, VA: Association for Supervision and Curriculum Development.

Popham, W. J. (2008). *Transformative assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

Powers, J. M. (2004). High-stakes accountability and equity: Using evidence from California's public schools accountability act to address the issues in *Williams v. State of California*. *American Educational Research Journal, 41*(4), 763-795.

Pyrczak, F. (2006). *Making sense of statistics*. Glendale, CA: Pyrczak Publishing.

Raymond, M. E., & Hanushek, E. A. (2003, Summer). High-stakes research. *Education Next, 3*(3), 48-55. Retrieved March 1, 2010, from www.educationnext.org

Reinard, J.C. (2006). *Communication research statistics*. Thousand Oaks, CA: Sage Publications, Inc.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools and academic achievement. *Econometrica, 73*, 417–458.

Robelen, E. W. (2000, May). Louisiana set to retain 4th, 8th graders based on state exams. *Education Week,* p. 24.

Roby, D. E. (2004). Research on school attendance and student achievement. A study of Ohio schools. *Educational Researcher Quarterly, 28*(1), 3-14.

Rogers, J. (2006). Forces of accountability? The power of poor parents in NCLB. *Harvard Educational Review*, *76*(4), 611-641. Retrieved January 3, 2010, from Academic Search Premier database.

Rosenshine, B. (2003, August). High-stakes testing: Another analysis. *Education Policy Analysis Archives, 11*(24). Retrieved January 13, 2010, from http://epaa.asu.edu.epaa/v11n24/

Rothstein, R. (2009). Taking aim at testing. *American School Board Journal, 196*(3), 32-35. Retrieved January 24, 2010, from Academic Search Premier database.

Rowan, B., Correntti, R., & Miller, R. J. (2002). What large-scale survey research tells us about teacher effects on student achievement: Insights from the Prospects study of elementary schools. *Teachers College Record, 104*(8), 1525-1567.

Ruiz-Primo, M. (2005). *A multi-method and multi-source approach for studying fidelity of implementation.* Paper presented at the annual meeting of the American Education Research Association, Montreal, Canada.

Ryan, T.G. (2006). Performance assessment: Critics, criticism, and controversy. *International Journal of Testing,* 6 (1), 97-104. Retrieved August 16, 2009, from EBSCOhost database.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119-140.

Salomone, R.C. (2003). *Same, different, equal: Rethinking single-sex schooling.* New Haven: Yale University Press.

Schunk, D. H. (1996). Goal and self-evaluative influences during children's cognitive skill learning. *American Educational Research Journal, 33*, 359-382.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, and M. Scriven (Eds.), *Perspectives of curriculum evaluation, Volume I* (pp. 39-83). Chicago, IL: Rand McNally.

Sheldon, S. B. (2007). Improving student attendance with school, family, and community partnerships. *The Journal of Education Research, 100*(5), 267-275.

Sheldon, K.M., & Biddle, B.J. (1998). Standards, accountability, and school reform: Perils and pitfalls. *Teachers College Record, 100*(1), 164-180.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14.

Sloane, F.C., & Kelly, A. E. (2003). Issues in high-stakes testing programs. *Theory into Practice,* 42(1), 12-17. Retrieved August 10, 2009, from ProQuest database.

Sly, L. (1999). Practice tests as formative assessment improve student performance on computer-managed learning assessments. *Assessment and Evaluation in Higher Education, 24*(3), 339-343.

Smith, J. (2006). Examining the long-term impact of achievement loss during the transition to high school. *The Journal of Secondary Gifted Education, 17*(4), 211-221.

Smith, M. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practices, 10,* 7-11.

Smith, M. S.(1972). Equality of educational opportunity: The basic findings reconsidered. In F. Mosteller & D. P. Moynihan (Eds.), *On equality of educational opportunity* (pp. 230–342). New York: Vintage Books.

Smylie, M. A. (1988). The enhancement function of staff development: Organizational and psychological antecedents to individual teacher change. *American Educational Researcher Journal, 25*(1), 1-30.

Solley, B.. (2007). On standardized testing: An ACEI position paper. *Childhood Education*, 84(1), 31-37. Retrieved January 3, 2010, from Career and Technical Education. (Document ID: 1372934341).

Song, H. (2010). Critical Issues and common pitfall in designing and conducting impact studies in education: Lessons learned from the What Works Clearinghouse (Phase I). *Educational Evaluation and Policy Analysis, 32*(3), 351-371.

Spencer, B.D., & Wiley, D.E. (1981). The sense and the nonsense of school effectiveness. *Journal of Policy Analysis and Management, 1*(1), 43-52.

Stein, M. L., Berends, M., Fuchs, D., McMaster, K., Saenz, L., Yen, L., Fuchs, L. S., & Compton, D. L. (2008). Scaling Up an Early Reading Program: Relationships among teacher support, fidelity of implemntation, and student performance across different sites and years. *Educational Evaluation and Policy Analysis, 30*(4), 368-388.

Stouthamer-Loeber, M., & Loeber, R. (1988). The use of prediction data in understanding delinquency. *Behavioral Sciences and the Law, 6*(3), 333-354.

Stringfield, S. (1991). Introduction to the special issue on chapter 1 policy and evaluation. *Educational Evaluation and Policy Analysis, 12*(4), 325-327.

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan, 83*(10), 758-765.

Strauss, R. P., & Sawyer, E. A. (1986). Some new evidence on teacher and student competencies. *Economics of Education Review, 5*(1), 41-48.

Summers, A. A., & Wolfe, B. L. (1975). *Equality of educational opportunity quantified: A production function approach*. Philadelphia: Federal Reserve Bank of Philadelphia.

Taylor, F. (1911). *The principles of scientific management.* New York and London: Harper & Brothers Publishers.

Taylor, J. K., & Dale, R. (1971). *A survey of teachers in the first year of service.* Bristol: University of Bristol, Institute of Education.

TeacherPortal (2009). New Jersey school districts. Retrieved August 26, 2009, from http://teacherportal.com/district/new-jersey

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach.* New York: The Guilfold Press.

Tienken, C.H. (2008a). A descriptive study of the technical characteristics of the results from New Jersey's assessments of skills and knowledge in grades 3,4, and 8. *New Jersey Journal of Supervision and Curriculum Development*, *52*, 46-61.

Tienken, C.H. (2008b). The characteristics of state assessment results. *Academic Exchange Quarterly, 12*(3), 34-39.

Tienken, C.H. & Rodriguez, O. (2010). The error of state mandated high school exams. *Academic Exchange Quarterly, 14*(2), 50-55.

Thompson, B. (2006). Foundations of behavioral statistics: An insight-based approach. New York: The Guilfold Press.

United States Census Bureau (2008). American Factfinder. Retrieved August 26, 2009, from http://factfinder.census.gov/home/saff/main.html?_lang=en

United States Department of Education. (1983). *A nation at risk: The imperative for educational reform* [On-line]. Available: http://www.ed.gov/pubs/NatATRisk/index.html

United States Department of Education. (2002). *No Child Left Behind: A desktop reference.* Jessup, MD: Education Publications Center.

United States Department of Education. (2009). *A highly qualified teacher in every classroom: The Secretary's sixth annual report on teacher quality.* Retrieved February 10, 2010, from http://www2.ed.gov/about/reports/annual/teachprep/index.html

Velan, G. M., Rakesh, K. K., Mark, D., & Wakefield, D. (2002). Web-based self-assessments in Pathology with Questionmark Perception. *Pathology, 34*, 282-284.

Vornberg, J., & Hart, R. (2000). *Accountability and high stakes testing: Views from Texas schools.* Paper presented at National Council of Professors of Educational Administration conference, Ypsilanti, MI.

Wang, T. H. (2007). What strategies are effective for formative assessment in an e-learning environment? *Journal of Computer Assisted Learning, 23*, 171-186.

Whitaker, J. (1997). *Interpretation of structure coefficients can prevent erroneous conclusions about regression results.* (ERIC Document No. ED 406438)

White, B.Y., & Frederiksen, J.R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction, 16*(1), 3-118.

Whiting, B., Van Burgh, J. W., & Render, G F. (1995). *Mastery learning in the classroom.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Wilkins, J. L. M., Zembylas, M., & Travers, K. J. (2002). Investigating correlates of mathematics and science literacy in the final year of secondary school. In D. F. Robitaille & A. E. Beaton (Eds.), *Secondary analysis of the TIMSS data* (pp. 291-316). Boston, MA: Kluwer Academic Publishers.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment.* Mahwah, NJ: Lawrence Erlbaum Associates.

Wininger, R. S. (2005). Using your tests to teach: Formative summative assessment. *Teaching Psychology, 32*(2), 164-166.

Woodbridge Township School District (2008-2009). *Middle school program of studies guide.* Retrieved March 15, 2010, from http://www.woodbridge.k12.nj.us/SchoolsMS/ms_pos_1011.pdf

Zhao, Y. (2009). *Catching up or leading the way: American education in the age of globalization.* Alexandria, VA: ASCD.